

Über zweistufige quadratische Klassifikatoren

Dem Fachbereich Ingenieurwissenschaften der
Universität Erlangen-Nürnberg
zur Erlangung des Grades
DOKTOR-INGENIEUR

vorgelegt von
Rainer Ott

Erlangen, 1977



Digitized by the Internet Archive
in 2026

https://archive.org/details/IA41548420_0085

Über zweistufige quadratische Klassifikatoren

Dem Fachbereich Ingenieurwissenschaften der
Universität Erlangen-Nürnberg
zur Erlangung des Grades
DOKTOR-INGENIEUR

vorgelegt von
Rainer Ott

Erlangen, 1977

Als Dissertation genehmigt vom
Fachbereich Ingenieurwissenschaften
der Universität Erlangen-Nürnberg

Tag der Einreichung: 4. Mai 1977
Tag der Promotion : 22. Juli 1977
Dekan : Prof. Dr. H. J. Schneider
Berichterstatter : Prof. Dr. H. Niemann
Prof. Dr. G. Meyer-Brötz

<u>Inhalt</u>	<u>Seite</u>
Zusammenstellung der wichtigsten Formelzeichen und Abkürzungen in der Reihenfolge ihres Auf- tretens	III
1. Einführung	1
2. Realisierung der 1. Stufe	8
2.1. Optimierung der Merkmalstransformation unter dem Gesichtspunkt der Rekonstruktion	11
2.2. Ordnen der Merkmale w_i unter dem Gesichts- punkt der Klassifizierung	
2.3. Experimenteller Vergleich der beiden Rang- ordnungen	28
2.4. Anschauliche Interpretationen zur Trans- formationsmatrix	37
3. Einblick in die Eigenschaften des zeichener- zeugenden Prozesses $\{v, y\}$ durch die Be- trachtung der Merkmale w_i	40
3.1. Eindimensionale Randverteilungen	45
3.2. Zweidimensionale Projektionen	50
3.3. Vergleich mit der M-dimensionalen Normal- verteilung	56
4. Realisierung der 2. Stufe	66
4.1. Normalverteilungs- und abstandsmessender Klassifikator	70
4.2. Quadratmittelklassifikator	74
4.3. Vergleichende Gegenüberstellung der qua- dratischen Klassifikatoren	82
5. Leistungsvergleich der betrachteten quadratischen Klassifikatoren am Beispiel der Handblockschrift- Stichprobe	94
5.1. Leistungsmessung an den Klassifikatoren nach 4.1.	98

5.2. Leistungsmessung an dem Klassifikator nach 4.2.	100
5.3. Gesichtspunkte zur praktischen Anwendung von Klassifikatoren nach 4.2.	108
5.4. Vergleich der Klassifikatoren nach 4.1. und 4.2. hinsichtlich der experimentellen Ergebnisse	114
6. Zusammenfassung	120
Literaturverzeichnis	126



Zusammenstellung der wichtigsten Formelzeichen und
Abkürzungen in der Reihenfolge ihres Auftretens

HAT	Abkürzung für Hauptachsentransformation
$\mathbf{V}$	Merkmalsvektor mit binär quantisierten Komponenten; mathematische Beschreibung der Rasterbilder eines Zeichens; Seite 6
$\mathbf{W}$	Merkmalsvektor mit reellen Komponenten; Seite 7; Gleichung (17)
N	Dimension des Merkmalsvektors $\mathbf{V}$; Zahl der Rasterbildelemente
M	Dimension des Merkmalsvektors $\mathbf{W}$; Zahl der für die Klassifikation benutzten Merkmale
$\mathbf{r}$	Fehlervektor. Er ergibt sich bei unvollständiger Rekonstruktion eines Zeichens; Gleichung (1) und (2)
R^2	Rekonstruktionsfehler; gemittelttes Quadrat des Fehlervektors; Gleichung (1)
$\mathbf{V}^*$	Merkmalsvektor, der durch Rücktransformation des M-dimensionalen Vektors $\mathbf{W}$ in das ursprüngliche N-dimensionale Koordinatensystem entsteht; Gleichung (3)
$\mathbf{B}$	Transformationsmatrix; die Spaltenvektoren sind Eigenvektoren von $\mathbf{K}$; Gleichung (4)
$\mathbf{b}_i$	Spaltenvektoren von $\mathbf{B}$; Eigenvektoren von $\mathbf{K}$; Gleichung (4)
$\mathbf{K}$	Kovarianzmatrix des Prozesses $\{\mathbf{v}\}$; Gleichung (21a)
$\mathbf{K}_k$	Kovarianzmatrix des Prozesses $\{\mathbf{w}/k\}$; Gleichung (45)
λ_i	Eigenwerte von $\mathbf{K}$; Gleichung (23)
$\mathbf{I}$	Einheitsmatrix
HDE	Abkürzung für Hauptdiagonalelement

K	Zahl der zu unterscheidenden Klassen
k	Klassenindex $k \leq K$
y	K-dimensionaler Zielvektor; diskrete Variable; Basisvektor eines K-dimensionalen orthonormalen Basisvektorsystems; Gleichung (33a)
d	Schätzvektor; Gleichung (33b)
d _k	k'te Komponente des Schätzvektors; Entscheidungsmaß; Schätzung der Rückschlußwahrscheinlichkeit.
A	Matrix mit K Spalten. Die Komponenten sind die Koeffizienten des Schätzpolynoms; Gleichung (33b), (92)
m _i	i'ter Zeilenvektor der Matrix E $\{wy^T\}$, Gleichung (38)
f	Fehlervektor der Entscheidung; Differenzvektor zwischen Schätzvektor d und dem Zielvektor y_{ent} der Entscheidungsklasse; Gleichung (115)
H0	Name der Handblockschrift-Stichprobe; Seite 28
A0	Name einer OCR-A Stichprobe (Nr. 0); Seite 29
A1	Name einer OCR-A Stichprobe (Nr. 1); Seite 29
A2	Name einer OCR-A Stichprobe (Nr. 2); Seite 29
LU	Abkürzung für <u>l</u> ineare <u>U</u> nabhängigkeit; Strategie zur Auswahl von Variablen; Seite 27
MS	Abkürzung für <u>m</u> aximale Rest <u>s</u> treuungsminderung, Strategie zur Auswahl von Variablen; Seite 27
NV	Abkürzung für Normalverteilung
Q _k	Quadratische Form, deren Kernmatrix die Inverse der klassenspezifischen Kovarianzmatrix ist; auch als Mahalanobis-Abstand bezeichnet; Gleichung (45)
L _k	Diagonalmatrix, deren Hauptdiagonalelemente Eigenwerte von K_k sind; Gleichung (49)
e	Klasse für die sich das Erkennungssystem entschieden hat; Gleichung (71)

R_i	Abkürzung für Risiko
r_k	Regressionsfunktion
S^2	Reststreuung; Gleichung (85)
X	Vektor, dessen Komponenten die Merkmale w_i und die Produkte $w_i w_j$ sind; Gleichung (93b)
DIF	Abkürzung für Differenz, Rückweisungskriterium; Seite 96
RAD	Abkürzung für Radius, Betrag des Fehlervektors f ; Rückweisungskriterium; Gleichung (116)
ξ	Rückweisungsrate
ζ	Fehlerrate
L	Länge des Schätzpolynoms; Gleichung (95) Hilfsvariable, Bedeutung ergibt sich aus Zusammenhang
FTN	Abkürzung für FORTRAN

1. Einführung

Durch den Einsatz moderner Rechenanlagen ist es möglich, sehr große Datenmengen innerhalb kurzer Zeit zu verarbeiten. Speziell dann, wenn die Daten von zahlreichen Bearbeitern dezentral erstellt wurden und anschließend zentral weiterverarbeitet werden sollen, stellt die Aufbereitung der Daten in unmittelbar maschinell lesbare Form zur Zeit noch einen ernstzunehmenden Engpaß dar. Die technische Entwicklung zielt daher in die Richtung, die menschliche Arbeitskraft, die zum Lesen und anschließenden Codieren der in Handschrift oder Maschinenschrift vorliegenden Dokumente eingesetzt wird, durch schnelle und zuverlässige Automaten weitgehend zu ersetzen. Diese Maschinen müssen, damit sie überhaupt und konkurrenzfähig eingesetzt werden können, mit hoher Lesesicherheit arbeiten.

Das Ziel der automatischen Erkennung von Schriftzeichen kann prinzipiell mit den verschiedensten Methoden erreicht werden. Es gibt eine kaum noch überschaubare Fülle von Vorschlägen für die Konstruktion automatisch arbeitender Erkennungssysteme. Nur ein sehr kleiner Anteil davon ist bis zur praktischen Realisierung ausgearbeitet oder wenigstens in die Nähe der praktischen Anwendbarkeit gebracht worden.

Trotzdem ist die Entwicklung auf dem Gebiet der automatischen Erkennungssysteme - und insbesondere dem der automatischen Schriftzeichenerkennung - nicht ungeordnet verlaufen. Die wissenschaftliche Diskussion hat auf diesem Arbeitsgebiet inzwischen zu klar gegliederten und überzeugenden Konzepten geführt, von denen für den Bereich der automatischen Schriftzeichenerkennung die aus der statistischen Betrachtungsweise hervorgehenden die weitaus größte Bedeutung haben. Es gibt inzwischen eine ganze Reihe einführender Fachbücher /A4, D1, F5, M3, M4, N3, N5, W4/, die diese Konzepte diskutieren. Ihr gemeinsamer Ausgangspunkt - der auch in der vorliegenden Arbeit benützt wird -, ist die Vorstellung, daß

die zu erkennenden Zeichen mit all den unvorhersehbaren Schwankungen ihres Erscheinungsbildes, Proben aus statistisch beschreibbaren Zufallsprozessen sind. Durch Messungen können die Eigenschaften der Zufallsprozesse erfaßt und für den Entscheidungsvorgang nutzbar gemacht werden. Ihrer Natur nach sind die dabei zu betrachtenden Zufallsvariablen viel-dimensional und erfordern zu ihrer Behandlung spezielle Verfahren der mathematischen Statistik. Die hier angeführten Überlegungen greifen dabei insbesondere auf $/A_1, L_1, R_1, R_2/$ zurück.

Im Hinblick auf die Konzeption von Erkennungssystemen hat sich die Gliederung in Meßdatengewinnung, Merkmalsbildung, Berechnung von Entscheidungsgrößen mit Hilfe von Unterscheidungsfunktionen und eigentliche Entscheidung durch Extremwertvergleich durchgesetzt. Dabei ist die Trennung zwischen Merkmalsbildung und Bildung der Unterscheidungsfunktion grundsätzlich problematisch. Ohne Beeinträchtigung der Erkennungsleistung sind Aufwandsverlagerungen von dem einen Bereich in den anderen möglich.

Die vorliegende Arbeit beschäftigt sich - bei vorgegebener Art und Weise der Meßdatenerfassung - mit den Teilsystemen Merkmalsbildung und Bildung der Unterscheidungsfunktionen, d. h. Klassifizierung. Dabei wird bei der Realisierung beider Teilsysteme Bezug genommen auf Lösungsmethoden, wie sie bereits von anderen Autoren vorgeschlagen wurden. Der Beitrag dieser Arbeit ist insbesondere darin zu sehen, daß der Bereich Merkmalsbildung und Klassifizierung unter den für die praktischen Anwendungen relevanten Bedingungen sorgfältig durchgearbeitet und diskutiert wird. Das betrifft einmal Art und Umfang des für die Untersuchungen verwendeten Stichprobenmaterials; zum anderen wird die Attraktivität der zugrunde gelegten Erkennungssystem-Konzeption für die praktische Verwendung in automatischen Lesemaschinen dargestellt. Diese Maschinen sollen imstande sein, die Forderungen nach hoher Verarbeitungsgeschwindigkeit und hoher Lesesicherheit

auch bei häufig schlechter Schrift- oder Druckqualität miteinander zu verbinden.

Im Bereich der Merkmalsbildung wird dabei der von verschiedenen Autoren diskutierten Möglichkeit gefolgt /A1, C1, F1, K3, N1, W1/, eine möglichst geringe Anzahl leistungsfähiger Merkmale durch Linearkombination der gegebenen Meßdaten zu gewinnen. Dieser Bereich kann durch die Stichworte Karhunen-Loeve Transformation bzw. Hauptachsentransformation (HAT) gekennzeichnet werden.

Die dabei erzeugten Merkmale sind nicht nur interessant als Eingabedaten für die nachfolgende Klassifizierung, sondern sie gestatten auch einen Rückblick auf die statistischen Eigenschaften der Meßdaten selbst. Daher bietet die statistische Analyse eben dieser Merkmale einen wichtigen Ansatzpunkt für die Interpretation der vorliegenden Zufallsprozesse und damit für das Verständnis der Aufgabenstellung. Aus derartigen Interpretationen gewonnene Kenntnisse haben erhebliche Bedeutung für die Konzeption des eigentlichen Erkennungssystems.

Im Bereich der Bildung der Unterscheidungsfunktionen wurden hauptsächlich zwei Konzepte diskutiert, die unter Benutzung verschiedener Annahmen und verschiedener Ausgangspunkte auf Unterscheidungsfunktionen des gleichen Typs führen: nämlich auf den in den Merkmalen vollständig quadratischen Klassifikator. Das erste Konzept /N1/ geht dabei von einer statistischen Hypothese über den Typ der klassenspezifischen Verteilung im Merkmalsraum aus. Benutzt wird hier die Annahme viel-dimensionaler Normalverteilungen mit verschiedenen Zentren und verschiedenen Kovarianzmatrizen. Dieser Ansatz führt auf quadratische Unterscheidungsfunktionen, deren Koeffizienten aus den statistischen Parametern Mittelwert und Kovarianz zu berechnen sind, die wiederum aus einer Lernstichprobe geschätzt wurden.

Das zweite Konzept /S1/ ist dagegen parameterfrei. Es werden keine Hypothesen über Verteilungstypen gemacht. Stattdessen werden die Koeffizienten der ebenfalls als vollständig quadratische Funktionen angesetzten Unterscheidungsfunktionen durch Quadratmitteladaption aus der Lernstichprobe berechnet.

Die angegebenen Stichworte kennzeichnen in Umrissen den Bereich, in dem die im folgenden dargestellten Überlegungen anzusiedeln sind. Um Zielsetzung und Beitrag dieser Arbeit deutlicher zu machen, ist es jedoch angebracht, die zugrunde gelegten Betrachtungsweisen ausführlich zu begründen.

Ausgangspunkt der Entwicklung eines Erkennungssystems ist der zeichenerzeugende Prozess. Dieser wird beschrieben durch eine möglichst umfangreiche und den wahren Prozess möglichst gut repräsentierende Stichprobe. In ihr sind Stichprobenelemente aller zu unterscheidenden Klassen enthalten.

Die verschiedenartig und verschieden häufig auftretenden Variationen von Zeichen aller vorkommenden Klassen kennzeichnen den Prozess. An das Erkennungssystem wird die Forderung gestellt, nicht alle beliebig denkbaren Formvariationen zu erkennen, sondern genau nur solche, wie sie überhaupt vorkommen. Dabei sind wiederum diejenigen Varianten von Zeichen, die häufiger auftreten, wichtiger als solche, die nur selten auftreten. Diese statistische Betrachtungsweise führt zu der entscheidungstheoretischen Lösung. Die möglichen Reaktionen des Erkennungssystems werden mit Kosten bewertet. Entwurfsziel ist die Minimierung des Risikos - des Erwartungswertes für die Kosten.

Die Arbeit geht konsequent vom entscheidungstheoretischen Lösungsansatz aus und behandelt dabei nur solche Methoden zur Entwicklung von Erkennungssystemen, die eine Adaption des Systems an gegebene Stichproben weitgehend automatisch ermöglichen. Unter Adaption wird dabei der einmal durchzuführende Lernprozess verstanden, in dem die frei verfügbaren

Parameter des Systems eingestellt werden. Die Vorgabe von Methode und gestellter Aufgabe - definiert durch eine in Klassenzahl, Zeichenanzahl, Schriftart und Schriftqualität variierende Stichprobe - soll die vollständige Realisierung des Systems ermöglichen, ohne daß spezielle Eingriffe durch den Menschen an wesentlichen Stellen erforderlich sind. Betrachtet wird ausschließlich die Menge der Einzelzeichen-erkennungssysteme, bei denen die Entscheidung für eine Bedeutungsklasse unabhängig von der Stellung dieses Zeichens in der vorliegenden Zeichenfolge gefällt wird.

Wichtiges Hilfsmittel während der Adaptionsphase eines Erkennungssystems an eine Stichprobe ist die Rechenanlage. Die Adaptionsfähigkeit des Systems liegt dabei in den Koeffizienten der Schätzfunktionen, die bei gegebener Struktur variiert werden dürfen.

Während der Adaptionsphase werden diese Klassifikatorkoeffizienten optimal an den zeichenerzeugenden Zufallsprozess angepaßt. Die heutigen zur Verfügung stehenden Anlagen, die sich durch hohe Verarbeitungsgeschwindigkeit und durch großen Arbeitsspeicher mit kurzer Zugriffszeit auszeichnen, ermöglichen während dieser Lernphase die Anwendung auch rechenintensiver Algorithmen. Damit ist in vielen Fällen überhaupt erst die Möglichkeit gegeben, theoretisch effektiv erscheinende Methoden auch praktisch zu realisieren. Die Leistungen von Erkennungssystemen, die nach verschiedensten Techniken entwickelt wurden, können durch Simulation auf der Rechenanlage unmittelbar verglichen werden. Dies bedeutet, diejenigen Funktionen durch Programm vorwegzunehmen, die später von speziell dafür eingesetzten elektronischen Bausteinen der Hardware übernommen werden. So ist bereits in einem frühen Zustand der Entwicklung eine Wertung verschiedener Systeme möglich. Ebenso kann durch Variation wichtig erscheinender Systemparameter deren grundsätzlicher Einfluß auf das Verhalten des Erkennungssystems verhältnismäßig leicht ausgemessen werden.

Die im praktischen Einsatz arbeitenden Lesemaschinen kann man als spezielle Prozessoren auffassen, die die Aufgabe haben, aus der aufbereiteten Information über das zu erkennende Zeichen und den in einem Speicher zur Verfügung stehenden optimierten Klassifikatordaten diejenigen Größen zu berechnen, die für die Entscheidung für eine Klasse notwendig sind. Geschwindigkeitserhöhung kann hier durch Parallelarbeit erreicht werden. Die rasche Entwicklung auf dem Gebiet der elektronischen Bauelemente (LSI-Bausteine, Mikroprozessoren) ermöglicht auch auf diesem Gebiet der Hardware den Einsatz von flexiblen Methoden.

Die moderne Rechnertechnologie stellt somit ein wirkungsvolles Handwerkzeug sowohl bei der Entwicklung wie bei der Realisierung von Systemkonzepten dar. Eine wichtige Aufgabe ist es nun, sich dieses Handwerkzeuges zu bedienen und, ausgehend von den verschiedensten theoretischen Ansätzen, Erkennungssysteme zu realisieren, deren spezielle Eigenschaften und Verhaltensweisen zu untersuchen und, wenn möglich, Systemkonzepte zu werten.

Die vorliegende Arbeit beschäftigt sich mit mehreren verschiedenen Methoden der Realisierung

zweistufiger quadratischer

Erkennungssysteme (Bild 1).

Es wird vorausgesetzt, daß auf die Bilder der abgetasteten Zeichen bereits ein wirkungsvoller Zeichenvorbereitungsalgorithmus, z. B. Zeilensuchen, Segmentieren, Bildreinigung, Zentrierung und Größennormierung, angewandt wurde. Ergebnis dieser Prozeduren ist das Rasterbild eines Zeichens, das N (etwa 200 bis 250) schwarzweiße Rasterbildelemente besitzt und mathematisch beschrieben werden kann durch einen N -dimensionalen Merkmalsvektor $\mathbf{V}$, der binär quantisierte Komponenten besitzt. ($\mathbf{V}$ wird häufig auch als Mustervektor bezeichnet).

Die erste Stufe des Systems hat die Aufgabe, die Information

über die Zeichen, dargestellt durch die N binär quantisierten Merkmale, nach Möglichkeit in wenig Dimensionen zu komprimieren, ohne daß dabei relevante Unterscheidungsmerkmale verloren gehen. Die erste Stufe erscheint damit als eine Merkmalsextraktionsstufe, die den Vektor $\mathbf{V}$ abbildet in einen Vektor $\mathbf{W}$, der nur noch M (M sehr viel kleiner als N) Komponenten besitzt.

Als zweite Stufe wird der Typus des in den Merkmalen w_i quadratischen Klassifikators betrachtet, der die für die eigentliche Entscheidung geeigneten Größen liefert.

Gegenstand der Arbeit ist das systematische Durcharbeiten verschiedener zweistufiger Klassifikator-Konzepte. Dadurch werden verschiedene Methoden, Möglichkeiten und Vorgehensweisen zur Realisierung von Erkennungssystemen nach wissenschaftlichen Gesichtspunkten geordnet. Es werden Bedingungen aufgezeigt, unter denen bestimmte Ansätze möglich sind und Ergebnisse angeführt, die sie rechtfertigen. Ziel der Arbeit ist es, im Bereich der Klassifikatorkonzeption und Klassifikatorentwicklung neue Konzepte zu erschließen und dabei die durch den Fortschritt der modernen Technologie greifbar gewordenen Möglichkeiten voll auszuschöpfen.

2. Realisierung der 1. Stufe

Die Aufgabe von automatischen Zeichenerkennungssystemen, aus den gemessenen Daten der zu klassifizierenden Zeichen Größen zu berechnen, die Maße für die Zugehörigkeit des Zeichens zu einer Klasse sind und aufgrund dieser Werte eine Entscheidung für eine Klasse zu treffen, läßt sich bei allen Systemen aufteilen in eine Reihe von Teilaufgaben mit genau definierten Funktionen. Die optoelektronische Wandlung der Information erfolgt mit dem Abtaster, ein Zeilensuchalgorithmus muß innerhalb einer vorgegebenen, je nach Aufgabe größeren oder kleineren Erwartungszone die richtige Zeile bestimmen, die gelesen werden soll, danach folgt die Segmentierung der Zeichenfolge in einzelne Zeichen. Die weiteren Prozeduren wie Bildreinigung, Zentrierung, Normierung sowie Abbildung auf ein gröberes Raster sind relativ leicht auszuführende Vorbereitungsschritte, die die Vielfalt der Erscheinungsformen von Zeichen mit gleicher Bedeutung einschränken und somit die noch auszuführende Klassifizierung wesentlich erleichtern. Diese Zeichenvorbereitungsprozeduren sind sehr wichtig und beeinflussen die Erkennungsleistung des gesamten Systems entscheidend.

Die beschriebenen Bildvorbereitungsprozeduren erzeugen ein Rasterbild mit ungefähr $N = 200$ bis 250 schwarzweiß Werten. Es kann durch den Merkmalsvektor $\mathbf{V}$ beschrieben werden, dessen N Komponenten binär quantisiert sind. Sämtliche Prozeduren zur Erzeugung dieses Merkmalsvektors werden in dieser Arbeit nicht betrachtet. Sie werden in /S2, S3, M1, G2/ ausführlich beschrieben und gewertet. Die Arbeit geht unmittelbar von dem Ergebnis der Vorbereitungsprozeduren, dem Merkmalsvektor $\mathbf{V}$ aus. Er ist Eingabegröße der 1. Stufe der hier betrachteten Erkennungssysteme. Diese Stufe wird in dieser Arbeit als Merkmalsextraktionsstufe definiert. Sie hat die Aufgabe, aus dem Merkmalsvektor $\mathbf{V}$ möglichst wenige, aber für die Klassenzugehörigkeit des vorliegenden Zeichens möglichst wichtige Merkmale zu extrahieren. Sämtliche M erzeug-

ten Merkmale ($M \approx 10 \dots 40$) werden im Merkmalsvektor $\mathbf{w}$ zusammengefaßt. Die 2. Stufe des Erkennungssystems ist ein nichtlinearer Klassifikator, der ausschließlich aufgrund der eingegebenen Merkmale w_i Maße für die Zugehörigkeit zu den verschiedenen Klassen bestimmt. Die Entscheidung für eine Klasse wird danach nur durch Absuchen des besten Maßes unter allen Klassen gefällt (Maximum- bzw. Minimumsuche).

Diese hier definierte Aufteilung in Merkmalsextraktions- und Klassifikationsstufe ist nur dann sinnvoll, wenn der Umfang der auszuführenden Aufgaben beider Stufen in einem vernünftigen Verhältnis zueinander steht. Es wird gefordert, daß weder der Aufgabenumfang der 1. Stufe gegenüber der 2. Stufe vernachlässigt werden kann, - denn dann übernimmt die Klassifikationsstufe praktisch die Arbeit der Merkmalsextraktionsstufe - noch umgekehrt, dann würden Merkmale von der ersten Stufe erzeugt werden, die bereits als Entscheidungsgrößen aufgefaßt werden können.

Das Problem der Erzeugung von Merkmalen, die für die zu erkennenden Muster signifikant sind, sowie das Bewerten dieser Merkmale nach ausgesuchten Kriterien, um die wirkungsvollsten unter den zur Diskussion stehenden Merkmalen zu bestimmen, sind Gegenstand vieler Veröffentlichungen (Engl. Stichworte: Feature Extraction, Feature Ordering, Feature Ranking) /A1, A2, B1, C1, C2, C3, F1, F2, F3, H1, K2, K3, N1, N2, N3, N4, O1, R1, T2, T3, T4, W1, W2, W3/. In /N3/ werden die wichtigsten Methoden in einer zusammenfassenden Darstellung beschrieben.

Aus vielen Literaturstellen geht deutlich hervor, daß die Bewertung der erzeugten Merkmale hinsichtlich der Wichtigkeit für die Klassifizierung mittels zusätzlich eingeführter Kriterien letztlich nur ein behelfsmäßiges Vorgehen ist, um von vornherein kaum oder wenig relevant scheinende von wichtiger Information zu trennen. Eine endgültige Aussage dar-

über, inwieweit geeignete Merkmale bestimmt wurden, kann erst dann gemacht werden, wenn die Leistung des gesamten Erkennungssystems - bestehend aus Merkmalsextraktions- und Klassifikationsstufe - bezüglich einer möglichst umfangreichen Referenzstichprobe ermittelt wird. Bei gleicher Aufgabenstellung sind dabei Rückweisungs- und Fehlerraten als Leistungsmaße objektive Bewertungsgrößen. Diese Bewertungsgrößen sind natürlich abhängig davon, wie wirkungsvoll der Klassifikationsalgorithmus ist, der schließlich die richtige oder falsche Entscheidung trifft. So kann eine erfolversprechende Merkmalsgewinnung durch einen zu einfachen Klassifizierungsalgorithmus nicht zum Tragen kommen. Dies gilt auch für den umgekehrten Fall, in dem ein wirkungsvoller Klassifizierungsalgorithmus mit einfacher Merkmalsgewinnung letztlich zu einem guten Erkennungsergebnis führt. Es bleibt festzustellen, daß voneinander unabhängiges Bewerten von Merkmalsextraktor und Klassifikator kaum möglich ist. Objektiver Beurteilung zugänglich jedoch ist das effektive Zusammenwirken beider Einheiten, die sich in der Leistung des Erkennungssystems ausdrückt.

In dieser Arbeit wird eine Merkmalsextraktionsstufe betrachtet, die aus dem N-dimensionalen Merkmalsvektor $\mathbf{v}$ durch eine lineare Orthogonaltransformation einen Vektor $\mathbf{w}$ erzeugt. Der zeichenerzeugende Zufallsprozess $\{\mathbf{v}\}$ wird somit abgebildet in einen Prozess $\{\mathbf{w}\}$, wobei jeder einzelne Merkmalsvektor $\mathbf{w}$ als eine Beschreibung für den ursprünglichen Merkmalsvektor $\mathbf{v}$ aufgefaßt werden kann, aus dem er hervorgegangen ist.

Vollständig fehlerfreie Beschreibung bedeutet i. a. die Verwendung von wiederum N Merkmalen w_i . Um die gewünschte Informationskompression zu erhalten, wird auf diese vollständige Beschreibung des Prozesses $\{\mathbf{v}\}$ verzichtet. Von den potentiellen N in Frage kommenden Merkmalen w_i werden nur $M \ll N$ verwendet. Der N-dimensionale Merkmalsraum wird somit abgebildet in einen M-dimensionalen Unterraum. Dabei

wird das Ziel verfolgt, den sich zwangsläufig ergebenden Fehler in der Beschreibung des Prozesses $\{\mathbf{v}\}$ überhaupt klein zu halten und dafür zu sorgen, daß die weggelassene Information möglichst wenig wichtig für die Unterscheidung der Zeichenklassen ist.

Wird jedes Zeichen durch einen Punkt im N-dimensionalen Merkmalsraum dargestellt, der sich an der Spitze des zugehörigen Merkmalsvektors $\mathbf{v}$ befindet, so erscheint der gesamte Zufallsprozess als Punktwolke im N-dimensionalen Merkmalsraum. Die lineare Abbildung in einen M-dimensionalen Unterraum kann nun aufgefaßt werden als ein erneutes Ausmessen der Lage jedes Punktes in einem neuen Koordinatensystem, das gegenüber dem ursprünglichen System gedreht und verschoben ist. Durch geeignete Wahl von Koordinatenmittelpunkt und Drehwinkel kann die Dimension M des neuen Systems klein gehalten werden, ohne dabei einen erheblichen Fehler in der Beschreibung des Prozesses $\{\mathbf{v}\}$ in Kauf nehmen zu müssen. Insbesondere ist eine drastische Dimensionsreduktion dann möglich, wenn die Ausdehnung der Punktwolke in den verschiedenen Raumrichtungen stark unterschiedlich ist. Deutlich wird dabei die unmittelbare Abhängigkeit dieses Beschreibungsfehlers von den statistischen Eigenschaften des Zufallsprozesses, d. h., der Fehler ist bei verschiedenen Stichproben (etwa hand- und maschinengeschriebenen Zeichen) und gleicher Transformationsvorschrift im allgemeinen unterschiedlich.

2.1. Optimierung der Merkmalstransformation unter dem Gesichtspunkt der Rekonstruktion

Ausgangspunkt bei der Dimensionierung der Merkmalsextraktionsstufe ist die Argumentation, daß der mathematisch definierte Rekonstruktionsfehler R^2 ein geeignetes Maß dafür ist, wie gut die Beschreibung des N-dimensionalen Zufallsprozesses $\{\mathbf{v}\}$ durch den Prozess $\{\mathbf{w}\}$ mit vorgegebener Zahl M von

Dimensionen gelungen ist. Um diese Argumentation deutlich werden zu lassen, werden im folgenden die an sich bekannten Zusammenhänge noch einmal dargestellt /A1, M3, W1/. Der Rekonstruktionsfehler R^2 wird definiert als über den gesamten Prozess gemitteltes Quadrat des Fehlervektors

$$R^2 = E \{ |\tau|^2 \} \quad (1)$$

τ ist dabei der Differenzvektor zwischen dem Merkmalsvektor V und dem Vektor V^* , der sich durch Rücktransformation des Vektors W in den N-dimensionalen Merkmalsraum ergibt.

$$\tau = V - V^* \quad (2)$$

V^* ist somit die mathematische Beschreibung des aus W rekonstruierten Rasterbildes.

Unter der einschränkenden Bedingung, nur lineare Abbildungen zuzulassen, ist es nun ohne größere Schwierigkeiten möglich, diejenige Abbildung zu bestimmen, die zu einem minimalen Rekonstruktionsfehler bei gegebener Zahl von Dimensionen führt

$$R^2 = \text{Minimum bei vorgegebenem } M.$$

Bei diesem Ansatz werden beide Stufen des Erkennungssystems als völlig entkoppelt betrachtet, d. h., von der Klassifikationsstufe wird keine Information an die Merkmalsextraktionsstufe rückgekoppelt. Die Dimensionierung der 1. Stufe erfolgt mit den eingeführten Hilfoptimierungskriterien (minimaler Rekonstruktionsfehler) völlig unabhängig von dem Aufbau der 2. Stufe. Diese Entkoppelung entspricht einer Aufteilung eines komplexen Problems in zwei überschaubare Teilprobleme. Man hofft, durch optimale unabhängige Lösung der beiden Teilprobleme auch eine gute Lösung des gesamten Problems zu erreichen, d. h., ein Erkennungssystem hoher Leistung zu erhalten.

Um den in das ursprüngliche Koordinatensystem rücktransformierten Vektor $\mathbf{v}^*$ aus dem M-dimensionalen Vektor $\mathbf{w} = (w_1 \dots w_M)^T$ zu bestimmen, wird folgender Ansatz gemacht:

$$\mathbf{v}^* = \sum_{i=1}^M w_i \mathbf{b}_i + \mathbf{c} \quad (3)$$

Entsprechend Gleichung (3) wird somit $\mathbf{v}^*$ als Summe mit M Gliedern geschrieben. Die vollständige fehlerfreie Darstellung $\mathbf{v} = \mathbf{v}^*$ ergibt sich als Sonderfall bei $\text{Dim } \{\mathbf{v}\} = N = M = \text{Dim } \{\mathbf{w}\}$. Faßt man die Vektoren $\mathbf{b}_i$ als Spaltenvektoren der Matrix $\mathbf{B}$ auf

$$\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2 \dots \mathbf{b}_M) \quad (4)$$

so kann Gleichung (3) auch geschrieben werden:

$$\mathbf{v}^* = \mathbf{B}\mathbf{w} + \mathbf{c} \quad (5)$$

Damit ergibt sich der Rekonstruktionsfehler

$$R^2 = E\{|\boldsymbol{\tau}|^2\} = E\{|\mathbf{v} - \mathbf{v}^*|^2\} = E\{|\mathbf{v} - \mathbf{c} - \mathbf{B}\mathbf{w}|^2\} \quad (6a)$$

Das Optimierungsziel ist demnach genau definiert durch

$$R^2 = \min_{\mathbf{B}} \left[\min_{\mathbf{c}} \left(\min_{\mathbf{w}} E\{|\boldsymbol{\tau}|^2\} \right) \right] \quad (6b)$$

Als Nebenbedingung fordern wir die Orthogonalität von $\mathbf{B}$

$$\mathbf{B}^T \mathbf{B} = \mathbf{I} \quad (7)$$

wobei $\mathbf{I}$ die M-dimensionale Einheitsmatrix ist. Ebenso soll die Triviallösung, bei der der Rekonstruktionsfehler für alle M verschwindet, ausgeschlossen sein. Gleichung (6b) ist eine Extremwertaufgabe hinsichtlich $\mathbf{c}$ und $\mathbf{B}$. Diese Größen werden unter Berücksichtigung des Erwartungswertoperators bestimmt ($E\{|\boldsymbol{\tau}|^2\} = \min$). Hinsichtlich $\mathbf{w}$ geschieht die Optimierung

für jedes einzelne Stichprobenelement $\mathbf{v}$

$$(|\mathbf{r}|^2 = \min_{\mathbf{w}} \text{ für alle } \mathbf{v}).$$

Mit der Abkürzung

$$\mathbf{v}_1 = \mathbf{v} - \mathbf{c} \quad (8)$$

ergibt sich das Betragsquadrat des Fehlervektors

$$|\mathbf{r}|^2 = \mathbf{r}^T \mathbf{r} = (\mathbf{v}_1 - \mathbf{B}\mathbf{w})^T (\mathbf{v}_1 - \mathbf{B}\mathbf{w}) \quad (9a)$$

$$= \mathbf{v}_1^T \mathbf{v}_1 - 2 \mathbf{v}_1^T \mathbf{B}\mathbf{w} + \mathbf{w}^T \mathbf{B}^T \mathbf{B}\mathbf{w} \quad (9b)$$

$$= \mathbf{v}^T \mathbf{v} - 2 \mathbf{c}^T \mathbf{v} + \mathbf{c}^T \mathbf{c} - 2 (\mathbf{v} - \mathbf{c})^T \mathbf{B}\mathbf{w} + \mathbf{w}^T \mathbf{w} \quad (9c)$$

Um die durch Gleichung (6b) definierte Extremwertaufgabe zu lösen, wird partiell nach $\mathbf{w}$, nach $\mathbf{c}$, und schließlich nach $\mathbf{B}$ abgeleitet (Symbole $\nabla_{\mathbf{w}}$, $\nabla_{\mathbf{c}}$, $\nabla_{\mathbf{B}}$). Durch Nullsetzen der Ableitungen ergeben sich die Bestimmungsgleichungen.

a) Partielle Ableitung nach $\mathbf{w}$ zur Bestimmung der optimalen Koeffizienten w_i für jedes $\mathbf{v}$ und bei konstantem $\mathbf{B}$ und $\mathbf{c}$.

$$\nabla_{\mathbf{w}} R^2 = \nabla E\{|\mathbf{r}|^2\} = E\{\nabla(\mathbf{v}_1^T \mathbf{v}_1 - 2 \mathbf{v}_1^T \mathbf{B}\mathbf{w} + \mathbf{w}^T \mathbf{w})\} \quad (10a)$$

$$= E\{\nabla(\mathbf{w}^T \mathbf{w} - 2 \mathbf{w}^T \mathbf{B}^T \mathbf{v}_1)\} \quad (10b)$$

$$= 2 E\{\mathbf{w} - \mathbf{B}^T \mathbf{v}_1\} \quad (10c)$$

Durch Nullsetzen der Ableitung unter Ausklammerung des Erwartungswertoperators erhalten wir die Gleichung

$$\mathbf{w} = \mathbf{B}^T \mathbf{v}_1 = \mathbf{B}^T (\mathbf{v} - \mathbf{c}) \quad (11)$$

Somit erhalten wir den Fehlervektor nach Gleichung (11), (5), (2):

$$\begin{aligned}\mathbf{r} &= \mathbf{v} - \mathbf{v}^* = \mathbf{v} - (\mathbf{B}\mathbf{B}^T(\mathbf{v} - \mathbf{c})) - \mathbf{c} \\ &= \mathbf{v}_1 - \mathbf{B}\mathbf{B}^T\mathbf{v}_1 = (\mathbf{I} - \mathbf{B}\mathbf{B}^T)\mathbf{v}_1\end{aligned}\quad (12)$$

Es wird unmittelbar deutlich, daß für den Sonderfall gleicher Dimension der Vektoren $\mathbf{v}$ und $\mathbf{w}$ nämlich $M = N$ der Fehlervektor $\mathbf{r}$ stets verschwindet. In diesem Fall ist $\mathbf{B}\mathbf{B}^T = \mathbf{I}$ und somit nach Gleichung (12) $\mathbf{r} = \mathbf{0}$. Für jeden nicht verschwindenden Fehlervektor $\mathbf{r}$ ($M < N$) gilt allerdings wegen Gleichung (7):

$$\mathbf{B}^T\mathbf{r} = (\mathbf{B}^T - \mathbf{B}^T \cdot \mathbf{B}\mathbf{B}^T)\mathbf{v}_1 = \mathbf{0} \quad (13)$$

Dies bedeutet, daß der Fehlervektor $\mathbf{r}$ eines beliebigen Merkmalsvektor $\mathbf{v}$ senkrecht zu allen übrigen $M-N$ Basisvektoren $\mathbf{b}_i$ ist.

b) Partielle Ableitung nach $\mathbf{c}$

Mit Gleichung (12) erhalten wir den Rekonstruktionsfehler R^2

$$\begin{aligned}R^2 &= E\{\mathbf{r}^T\mathbf{r}\} = E\{\mathbf{v}_1^T(\mathbf{I} - \mathbf{B}\mathbf{B}^T)^T(\mathbf{I} - \mathbf{B}\mathbf{B}^T)\mathbf{v}_1\} \\ &= E\{\mathbf{v}_1^T(\mathbf{I} - \mathbf{B}\mathbf{B}^T)\mathbf{v}_1\}\end{aligned}\quad (14)$$

Nun ist nur die Variable $\mathbf{v}_1$ von $\mathbf{c}$ abhängig. Damit ergibt sich:

$$\begin{aligned}\nabla_{\mathbf{c}} R^2 &= -\nabla_{\mathbf{v}_1} R^2 = -2 E\{(\mathbf{I} - \mathbf{B}\mathbf{B}^T)\mathbf{v}_1\} \\ &= -2 (\mathbf{I} - \mathbf{B}\mathbf{B}^T) \cdot (E\{\mathbf{v}\} - \mathbf{c})\end{aligned}\quad (15)$$

Der erste Ausdruck $(\mathbf{I} - \mathbf{B}\mathbf{B}^T)$ ist für $M < N$ stets ungleich $\mathbf{0}$. Deswegen ergibt sich das Extremum durch Nullsetzen des 2. Faktors

$$C = E\{v\} = \bar{v} \quad (16)$$

Somit ist die bisher unbekannte Konstante C derart bestimmt worden, daß R^2 bezüglich C unabhängig von der Matrix B zu einem Extremum wird. Dieses Extremum ist in dem vorliegenden Fall ein Minimum, da R^2 als definitionsgemäß nicht negative Größe den Wert ∞ für $C \rightarrow \infty$ besitzt. Daher genügt es, das Vorhandensein verschwindender Ableitungen zu zeigen.

c) Optimale Wahl des Basisvektorsystems

Nach der Bestimmung der Transformationsvorschrift zur Berechnung von w , sowie der optimalen Wahl von C müssen noch die Basisvektoren b_i ermittelt werden. Nach Gleichung (11) und (16) gilt nun:

$$w = B^T(v - \bar{v}) \quad (17)$$

Der zu minimierende Rekonstruktionsfehler ergibt sich entsprechend Gleichung (14) zu

$$R^2 = E\{v_1^T v_1\} - E\{v_1^T B B^T v_1\} \quad (18)$$

Definieren wir K als die Kovarianzmatrix des Zufallsprozesses $\{v\}$

$$K = E\{v_1 v_1^T\} = E\{(v - \bar{v})(v - \bar{v})^T\} \quad (19)$$

so läßt sich Gleichung (18) umschreiben

$$R^2 = E\{v_1^T v_1\} - \underset{B}{\text{Spur}\{B^T K B\}} \stackrel{!}{=} \underset{B}{\text{Min}} \quad (20)$$

Die Forderung nach Minimierung bezüglich $B = (b_1, \dots, b_M)$ zerfällt in eine Summe von Teilforderungen

$$R^2 = E\{\mathbf{v}_1^T \mathbf{v}_1\} - \sum_{i=1}^M \mathbf{b}_i^T \mathbf{K} \mathbf{b}_i = \min_{\mathbf{b}_i} \quad (21a)$$

$$= E\{\mathbf{v}_1^T \mathbf{v}_1\} - \sum_{i=1}^M Q_i \quad (21b)$$

Q_i ist eine quadratische Form mit einer Kovarianzmatrix als Kernmatrix und daher gilt:

$$Q_i \geq 0 \quad (22)$$

Demnach wird der Rekonstruktionsfehler R^2 nach Gleichung (21b) zu einem Minimum, wenn alle Q_i maximal werden. Dies ist eine bekannte Extremwertaufgabe mit der Nebenbedingung

$$\mathbf{b}_i^T \mathbf{b}_i = 1$$

Unter Anwendung des Lagrange-Verfahrens bilden wir hierzu die Funktion

$$F(\mathbf{b}_i) = \mathbf{b}_i^T \mathbf{K} \mathbf{b}_i + \lambda_i (1 - \mathbf{b}_i^T \mathbf{b}_i) \quad (23)$$

Zur Bestimmung des Extremums wird nach $\mathbf{b}_i$ partiell differenziert und Null gesetzt

$$\nabla_{\mathbf{b}_i} F_i = 2 \mathbf{K} \mathbf{b}_i - 2 \lambda_i \mathbf{b}_i = 2(\mathbf{K} - \lambda_i \mathbf{I}) \mathbf{b}_i = 0 \quad (24)$$

Gleichung (24) definiert das bekannte Eigenwertproblem. Eine Lösung mit reellen, nicht negativen Werten λ_i gibt es, wenn $\mathbf{K}$ positiv definit ist. Da $\mathbf{K}$ als Kovarianzmatrix definiert ist, ist diese Bedingung erfüllt. Als Ergebnis der Aufgabe, den Rekonstruktionsfehler bezüglich $\mathbf{B}$ zu minimieren, erhalten wir die Vorschrift, M Eigenvektoren von $\mathbf{K}$ als Spaltenvektoren von $\mathbf{B}$ einzusetzen.

Die Ermittlung derjenigen M Eigenvektoren, die von allen N zur Verfügung stehenden Eigenvektoren von $\mathbf{K}$ als Spaltenvektoren von $\mathbf{B}$ einzusetzen sind, ergibt sich durch Berechnung des Rekonstruktionsfehlers R^2 bei M -dimensionalen Vektoren $\mathbf{w}$ nach Gleichung (21a).

$$R^2 = E\{\mathbf{v}_1^T \mathbf{v}_1\} - \sum_{i=1}^M Q_i = \text{Spur}\{\mathbf{K}\} - \sum_{i=1}^M Q_i \quad (25)$$

Dabei läßt sich Q_i in Abhängigkeit von den Eigenwerten direkt ausdrücken, indem Gleichung (24) von links mit $\mathbf{b}_i^T$ multipliziert wird.

$$0 = \mathbf{b}_i^T (\mathbf{K} - \lambda_i \mathbf{I}) \mathbf{b}_i = \mathbf{b}_i^T \mathbf{K} \mathbf{b}_i - \lambda_i = Q_i - \lambda_i \quad (26a)$$

$$\rightarrow Q_i = \lambda_i \quad (26b)$$

Somit ergibt sich R^2 zu

$$R^2 = \text{Spur}\{\mathbf{K}\} - \sum_{i=1}^M \lambda_i \quad (27)$$

Nun ist aber die Spur der N -dimensionalen Matrix gleich der Summe ihrer N Eigenwerte, womit sich R^2 vereinfacht.

$$\rightarrow R^2 = \sum_{i=M+1}^N \lambda_i \quad (28)$$

Damit liegt fest, daß beim Aufbau der Matrix $\mathbf{B}$ diejenigen M Eigenvektoren als Spaltenvektoren von $\mathbf{B}$ zu wählen sind, die zu den größten Eigenwerten gehören. Dann ist nach Gleichung (28) der Rekonstruktionsfehler bezüglich $\mathbf{B}$ minimiert.

Eine hervorzuhebende Eigenschaft der Transformation ist, nicht korrelierte Merkmale zu erzeugen. Um dies zu zeigen, bestimmen wir zunächst den Erwartungswert des Prozesses $\{\mathbf{w}\}$

$$\begin{aligned}\bar{\mathbf{w}} &= E\{\mathbf{w}\} = E\{\mathbf{B}^T(\mathbf{v}-\bar{\mathbf{v}})\} \\ &= \mathbf{B}^T(\bar{\mathbf{v}}-\bar{\mathbf{v}}) = \mathbf{0}\end{aligned}\quad (29)$$

und bilden nun die Kovarianzmatrix von $\{\mathbf{w}\}$

$$\begin{aligned}\mathbf{L} &= E\{\mathbf{w}\mathbf{w}^T\} = \mathbf{B}^T \cdot E\{(\mathbf{v}-\bar{\mathbf{v}})(\mathbf{v}-\bar{\mathbf{v}})^T\} \mathbf{B} \\ &= \mathbf{B}^T \mathbf{K} \mathbf{B}\end{aligned}\quad (30)$$

Um die Elemente der Kovarianzmatrix zu bestimmen, multiplizieren wir Gleichung (24) mit einem beliebigen Eigenvektor $\mathbf{b}_j$ von links und erhalten

$$\mathbf{b}_j^T (\mathbf{K} - \lambda_i \mathbf{I}) \cdot \mathbf{b}_i = 0 \quad (31a)$$

$$\mathbf{b}_j^T \mathbf{K} \mathbf{b}_i = \lambda_i \mathbf{b}_j^T \mathbf{b}_i \quad (31b)$$

Wegen der Orthogonalitätsbedingung der Eigenvektoren (Gleichung (7)) ergibt die rechte Seite von Gleichung (31b) stets 0, wenn $i \neq j$ ist, und 1, wenn $i = j$ ist

$$\mathbf{b}_j^T \cdot \mathbf{b}_i = \delta_{ji} \left\{ \begin{array}{l} = 0, i \neq j \\ = 1, i = j \end{array} \right\} \quad (31c)$$

Dies bedeutet, daß sämtliche Hauptdiagonalelemente (HDE) von $\mathbf{L}$ mit den Eigenwerten λ_i besetzt sind. Sämtliche Elemente außerhalb der Diagonale verschwinden jedoch:

$$E\{w_i w_i\} = \text{var}\{w_i\} = \lambda_i \quad (32a)$$

$$E\{w_i w_j\} = \text{cov}\{w_i, w_j\} = 0 \quad (32b)$$

Die folgende tabellarische Zusammenfassung der wichtigsten Eigenschaften der erzeugten Zufallsvariablen $\{\mathbf{w}\}$

soll abschließend das Wesentliche hervorheben.

• Transformationsvorschrift:

$$\mathbf{w} = \mathbf{B}^T(\mathbf{v} - \bar{\mathbf{v}})$$

- $\mathbf{B}$ ist eine rechteckige $N \times M$ Matrix ($M = \text{Dim}\{\mathbf{w}\}$, $N = \text{Dim}\{\mathbf{v}\}$). Zu ihrer Bestimmung wird zunächst die Kovarianzmatrix

$$\mathbf{K} = E\{(\mathbf{v} - \bar{\mathbf{v}})(\mathbf{v} - \bar{\mathbf{v}})^T\}$$

gebildet. Die M Eigenvektoren von $\mathbf{K}$

$$\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_M$$

die zu den M größten Eigenwerten

$$\lambda_1, \lambda_2, \dots, \lambda_M$$

gehören, sind die Spaltenvektoren von $\mathbf{B}$. Dabei gilt die Orthogonalitätsbedingung

$$\mathbf{B}^T \mathbf{B} = \mathbf{I}$$

- $\bar{\mathbf{v}}$ ist Schwerpunktsvektor des gesamten Zufallsprozesses $\{\mathbf{v}\}$.

$$\bar{\mathbf{v}} = E\{\mathbf{v}\}$$

• Rekonstruktionsfehler R^2

$$R^2 = \sum_{i=M+1}^N \lambda_i$$

• Eigenschaften des Prozesses $\{\mathbf{w}\}$

$$\bar{\mathbf{w}} = E\{\mathbf{w}\} = \mathbf{0}$$

Prozess ist zentriert

$$E\{\mathbf{w}\mathbf{w}^T\} = \mathbf{L}$$

Kovarianzmatrix ist
Diagonalmatrix mit
Eigenwerten λ_i als HDE

- $E\{w_i, w_j\} = \text{cov}\{w_i, w_j\} = 0; i \neq j$ Die Merkmale w_i sind nicht korreliert
- $E\{w_i, w_i\} = \text{var}\{w_i\} = \lambda_i$ Varianz von w_i ist Eigenwert λ_i
- * Merkmale w_i sind im Vektor $\mathbf{W}$ geordnet nach ihrer Varianz λ_i (diese Rangordnung wird mit LU bezeichnet - Abk. für Lineare Unabhängigkeit - siehe Kapitel 2.2.).

Die hergeleitete Abbildung wird als Hauptachsentransformation (HAT) bezeichnet. Die Hauptachsen wurden bestimmt an der Punktwolke im Merkmalsraum, die den stochastischen Prozess $\{\mathbf{v}\}$ als endliche Stichprobe repräsentiert. Die Stichprobe enthält dabei Merkmalsvektoren $\mathbf{V}$ aller zu unterscheidender Klassen. Zur Bestimmung der Transformationsmatrix $\mathbf{B}$ in der Lernphase wird hierbei die Kenntnis der Klassenzugehörigkeit eines jeden Rasterbildes nicht herangezogen. Der gesamte Prozess $\{\mathbf{v}\}$ wird bei dieser Abbildung als einheitlich aufgefaßt. Die jedem Merkmalsvektor zugehörige Kennung wird ausschließlich in der folgenden 2. Stufe, der Klassifikationsstufe, bei der Koeffizientenadaptation während des überwachten Lernens ausgewertet. Damit ist eine völlige Entkopplung der beiden Stufen des Erkennungsprozesses erreicht und daher ein unabhängiges Optimieren möglich: der Rekonstruktionsfehler wird ohne Kenntnis der Klassenzugehörigkeit minimiert.

In einigen Untersuchungen /G2, N1, N2/, in denen die HAT ebenfalls verwendet wird, wird im Gegensatz zu dieser Abbildung häufig zur Berechnung der Hauptachsen auf speziell ausgesuchte Teilprozesse Bezug genommen, die die Aufteilung des Merkmalsraumes in Klassengebiete voraussetzt. Damit wird auf die Eigenständigkeit der Stufen bewußt verzichtet.

Die HAT ist eng verwandt mit den verschiedenen diskreten Karhunen-Loeve-(KL)-Entwicklungen $/N_1/$, die wiederum ein Sonderfall der kontinuierlichen KL-Entwicklung sind. Diese entwickeln stochastische Prozesse nach orthogonalen Funktionssystemen. Ziel der Entwicklung ist, bei vorzeitigem Abbruch nach endlich vielen Gliedern einen minimalen Abbruchfehler zu erhalten. Dieser Abbruchfehler ist dabei zu vergleichen mit dem Rekonstruktionsfehler R^2 . Verschieden definierte Optimierungskriterien bei der KL-Entwicklung führen im wesentlichen dazu, die Eigenvektoren und Eigenwerte aus anderen, etwa klassenspezifisch gewichteten Kovarianzmatrizen bestimmen zu müssen.

2.2. Ordnen der Merkmale w_i unter dem Gesichtspunkt der Klassifizierung

Das Optimierungskriterium - Minimierung des Rekonstruktionsfehlers R^2 - zur Bestimmung der Merkmale w_i wurde ersatzweise eingeführt, um auf dem Wege zur Ermittlung der Klassenzugehörigkeit ein Zwischenziel zu erreichen. Die Optimierung kann dabei, wie bei vielen Arten der Merkmalsgewinnung, nicht direkt Bezug auf eine Klassifizierungsfähigkeit nehmen. Dies ist ausschließlich der 2. Stufe, der Klassifikationsstufe vorbehalten. Die Merkmalsgewinnung bezieht sich jedoch auf Kriterien, die in sehr enger Beziehung zur Klassifikation stehen. Man argumentiert dabei, daß es vernünftig ist, davon auszugehen, daß wenig ausgesuchte Merkmale, die eine gute Rekonstruktion der Zeichen ermöglichen, auch gut zur Klassifikation geeignet sind. Die so erzeugten Merkmale sind nun als gegeben anzusehen. Somit stehen alle N von der Hauptachsentransformation erzeugten Merkmale als potentielle Kandidaten für die Weiterverwendung in der Klassifikationsstufe zur Verfügung. Der für die Herleitung benutzte Gesichtspunkt der Rekonstruktion prägt den Merkmalen dabei eine natürliche Rangordnung auf, die Rangordnung nach der Größe der zugehörigen Eigenwerte λ_i .

Um aus diesen N Merkmalen eine Auswahl von wenigen zu bilden, kann man zunächst die am Anfang der natürlichen Rangordnung stehenden nehmen. Darüber hinaus gibt es natürlich die Möglichkeit, die mitgelieferte natürliche Rangordnung vollständig zu ignorieren und stattdessen eine neue zu bilden, eine Rangordnung, die sich jetzt am Begriff der Klassifizierung orientiert (Bild 2). Um eine neue Rangordnung auf diese Art aufzustellen, wird ein Brauchbarkeitsmaß benötigt, das in möglichst guter Übereinstimmung mit der Fähigkeit der Merkmale w_i zur Klassifizierung stehen sollte. Eine zusätzliche Bedingung, die an das Brauchbarkeitskriterium gestellt wird, ist die praktische Anwendbarkeit. Man möchte somit mit einem vertretbaren experimentellen Aufwand jedem einzelnen Merkmal ein Maß für die Klassifizierungsfähigkeit zuordnen, um so zu einer Rangordnung unter allen Merkmalen zu kommen. Die Methode, die zu der optimalen Lösung führt, besteht darin, alle bildbaren Untermengen mit M Merkmalen aus den N gegebenen mit einem daran adaptierten Klassifikator des vorgesehenen Typs (hier quadratischer Klassifikator) probeweise zu testen, und sich für diejenige zu entscheiden, die das beste Klassifizierungsergebnis liefert. Diese theoretische Lösung ist wegen der zu großen Zahl von Untermengen, die es zu testen gilt, praktisch nicht durchzuführen.

Um die an der Klassifizierungsfähigkeit orientierte Rangordnung zu ermitteln, prüfen wir, wie gut sich die wahre Klassenzugehörigkeit der Zeichen durch ein in den Merkmalen w_i lineares Polynom schätzen läßt /01/. Dazu ist es erforderlich, von einer den zeichenerzeugenden Prozess repräsentierenden Stichprobe auszugehen, die als Beschreibung jedes Zeichens den Merkmalssatz $w_i, i = 1 \dots M$ und die wahre Klassenzugehörigkeit - den Zielvektor y - enthält. Gegenstand der folgenden Untersuchung ist somit der Prozess $\{w, y\}$. Wir definieren y entsprechend /S1/ als einen Basisvektor in dem K -dimensionalen Entscheidungsraum (K ist die Zahl der zu unterscheidenden Klassen). Jedem Zeichen

der i'ten Klasse wird dabei der i'te Spaltenvektor der Einheitsmatrix **I** als Zielvektor **y** zugeordnet.

$$\begin{array}{c}
 \text{i'te Spalte} \\
 \downarrow \\
 \mathbf{I} = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ \vdots & \vdots & & 0 & \vdots \\ \vdots & \vdots & & 1 & \vdots \\ \vdots & \vdots & & 0 & \vdots \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & & 0 & 1 \end{bmatrix} \\
 \uparrow \\
 \text{Zielvektor } \mathbf{y} \text{ für Klasse } i
 \end{array} \quad (33a)$$

Der Vektor **y** ist damit als Zielvektor definiert, den es möglichst gut durch das Polynom

$$\mathbf{d} = \mathbf{A}_{Lin}^T \begin{pmatrix} 1 \\ \mathbf{w} \end{pmatrix} \quad (33b)$$

zu schätzen gilt. Die Koeffizienten der Matrix $\mathbf{A}_{Lin}$ werden so bestimmt, daß der mittlere quadratische Fehler S^2 zwischen Zielvektor **y** und Schätzvektor **d** zu einem Minimum wird

$$S^2 = \min_{\mathbf{A}_{Lin}} \left[E \{ |\mathbf{y} - \mathbf{d}(\mathbf{A}_{Lin})|^2 \} \right] \quad (34)$$

Dies bedeutet, den linearen Quadratmittelklassifikator zur Ermittlung der Rangreihenfolge unter den Merkmalen w_i heranzuziehen. Die Lösung der Extremwertaufgabe Gleichung (34) führt durch Ableitung von S^2 nach $\mathbf{A}_{Lin}$ /S1/ auf das Gleichungssystem

$$E \left\{ \begin{pmatrix} 1 \\ \mathbf{w} \end{pmatrix} \begin{pmatrix} 1, \mathbf{w}^T \end{pmatrix} \right\} \mathbf{A}_{Lin} = E \left\{ \begin{pmatrix} 1 \\ \mathbf{w} \end{pmatrix} \mathbf{y}^T \right\} \quad (35a)$$

$$E\left\{\begin{bmatrix} 1 & \mathbf{w}^T \\ \mathbf{w} & \mathbf{w}\mathbf{w}^T \end{bmatrix}\right\} \mathbf{A}_{Lin} = E\left\{\begin{bmatrix} \mathbf{y}^T \\ \mathbf{w}\mathbf{y}^T \end{bmatrix}\right\} \quad (35b)$$

Löst man Gleichung (35) nach $\mathbf{A}_{Lin}$ auf, und berücksichtigt die statistischen Eigenschaften des Prozesses $\{\mathbf{w}\}$ - zentrierter Prozess, Merkmale unkorreliert, durch Gleichung (29), (30), (32a, b) definiert - so ergibt sich

$$\begin{aligned} \mathbf{A}_{Lin} &= \begin{bmatrix} 1 & & & \mathbf{0}^T \\ & \lambda_1 & & \\ \mathbf{0} & & \ddots & \\ & & & \lambda_N \end{bmatrix}^{-1} \cdot \begin{bmatrix} \mathbf{y}^T \\ \overline{\mathbf{w}\mathbf{y}^T} \end{bmatrix} \\ &= \begin{bmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \mathbf{L}^{-1} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{y}^T \\ \overline{\mathbf{w}\mathbf{y}^T} \end{bmatrix} = \begin{bmatrix} \mathbf{y}^T \\ \mathbf{L}^{-1} \overline{\mathbf{w}\mathbf{y}^T} \end{bmatrix} \quad (36) \end{aligned}$$

Man sieht, daß in dem hier vorliegenden Fall nicht korrelierter Merkmale das Gleichungssystem in praktisch gelöster Form vorliegt, da die zu invertierende Matrix Diagonalform besitzt. Somit braucht nur jede Zeile $\mathbf{m}_i^T$ der Matrix $\overline{\mathbf{w}\mathbf{y}^T}$ durch das entsprechende Hauptdiagonalelement, den Eigenwert λ_i dividiert zu werden.

Zerlegt man die Koeffizientenmatrix $\mathbf{A}_{Lin}$ in ihre Zeilenvektoren $\mathbf{a}_i^T$

$$\mathbf{A}_{Lin} = \begin{bmatrix} \mathbf{a}_1^T \\ \mathbf{a}_2^T \\ \vdots \\ \mathbf{a}_N^T \end{bmatrix} \quad \text{mit} \quad \mathbf{a}_i^T = \frac{\mathbf{m}_i^T}{\lambda_i}; \quad i > 0 \quad (37)$$

dann wird bei der Berechnung des Schätzvektors

$$\mathbf{d} = \mathbf{A}_{Lin}^T \begin{pmatrix} 1 \\ \mathbf{w} \end{pmatrix} = \sum_{i=0}^N w_i \mathbf{a}_i \quad (33c)$$

jedem Merkmal w_i ein Vektor $\mathbf{a}_i$ zugeordnet. Dieser ergibt, multipliziert mit dem Wert der Variablen w_i , den Beitrag $\Delta \mathbf{d}_i$ des Merkmals zum Schätzvektor.

Der Vektor $\mathbf{d}$ wird, entsprechend Gleichung (33c), durch die Berücksichtigung aller Merkmale und ihrer zugehörigen Koeffizienten im Schätzpolynom berechnet. Damit wird ein über die Stichprobe gemittelter quadratischer Fehler S_N^2 erreicht. Nun interessieren wir uns aber gerade dafür, nur wenige M Variable w_i gezielt auszusuchen und in der Schätzgleichung zu berücksichtigen. Die zugehörigen Koeffizienten des Schätzpolynoms, die das Optimierungsziel Gleichung (34) erfüllen, sind in dem vorliegenden Fall nicht korrelierter Merkmale - und nur in diesem Fall - die gleichen Koeffizientenvektoren $\mathbf{a}_i$, wie bei voller Länge des Schätzpolynoms. Bei korrelierten Merkmalen w_i ergeben sich dagegen jeweils andere Koeffizientenvektoren, die als Zwischenlösung bei der Matrixinversion (schrittweise Regressionsanalyse) Gleichung (36) anfallen. Die Matrix

$$E \left\{ \begin{pmatrix} 1 \\ \mathbf{w}^* \end{pmatrix} \begin{pmatrix} 1 & \mathbf{w}^{*T} \end{pmatrix} \right\}$$

hat dabei keine Diagonalform.

Die Reststreuung S_M^2 , die mit dem Polynom der Länge M erreicht wird, ist dabei größer als der Endwert S_N^2 . Die Verlängerung der Schätzgleichung durch Berücksichtigung eines neuen Merkmals w_i erniedrigt die Reststreuung nach /S1, Gleichung (21)/ und Gleichung (37) um einen Betrag

$$\Delta S_i^2 = \frac{|m_i|^2}{\lambda_i} = |\mathbf{a}_i|^2 \cdot \lambda_i \quad (38)$$

Dieser Beitrag eines jeden Merkmals zur Minderung der Reststreuung ist bei unkorrelierten Merkmalen und wiederum nur

bei diesen, unabhängig von der Untermenge der Merkmale, die bereits zur Schätzung verwendet wurden. Der enge Zusammenhang zwischen Klassifikatorleistung und erreichter Reststreuung ist sicherlich unmittelbar ersichtlich (Definition von S^2). Darüber hinaus hat die langjährige Beschäftigung mit Quadratmittelklassifikatoren, bei denen die Argumente der Schätzpolynome die direkt gemessenen Merkmale v_i sind, gezeigt, daß die Einführung des Optimierungskriteriums S^2 vernünftig ist. Damit ist auch die Einführung von ΔS_i^2 als ein Maß zur Bewertung des Merkmals w_i hinsichtlich der Klassifizierungsfähigkeit hinreichend gerechtfertigt. Wir erhalten somit die neue Rangordnung unter den Merkmalen durch Ordnen ihrer Beiträge ΔS_i^2 . Wir bezeichnen diese als

MS-Rangordnung

/01/. Die ursprüngliche Rangordnung der Merkmale, die sich an der Rekonstruktionsfähigkeit orientiert und die von der Merkmalsextraktionsstufe unmittelbar als natürliche Rangordnung mitgeliefert wird, bezeichnen wir als

LU-Rangordnung

/01/. Beide Rangordnungen weisen eine formale Ähnlichkeit auf. Während bei der MS-Rangordnung jedem Merkmal w_i ein ΔS_i^2 zugeordnet wird, ist bei der LU-Rangordnung der Beitrag ΔR_i^2 zur Minderung des Rekonstruktionsfehlers entscheidend. Aus Gleichung (28) ist zu ersehen, daß dieser Beitrag der zugehörige Eigenwert ist

$$\Delta R_i^2 = \lambda_i = \text{var} \{w_i\} \quad (39)$$

Die folgende Tabelle verdeutlicht diesen formalen Zusammenhang beider Rangordnungen.

Kriterium	Ziel	Rangordnung unter den Merkmalen ergibt sich durch Ordnen der Größen	Eingeführte Bezeichnung
Rekonstruktionsfähigkeit	R^2 → Min	$\Delta R_i^2 = \lambda_i$	LU-Rangordnung
Klassifikationsfähigkeit	S^2 → Min	$\Delta S_i^2 = a_i ^2 \lambda_i$	MS-Rangordnung

Wir verfügen damit nun über zwei Rangordnungen. Dabei erhebt sich die Frage, ob und welche Unterschiede es zwischen beiden gibt. Die folgenden Untersuchungen sollen dies klären.

2.3. Experimenteller Vergleich der beiden Rangordnungen

In den durchgeführten Experimenten, die Aufschluß über das Verhalten der Merkmalsextraktionsstufe bezüglich realistischen Datenmaterials geben sollen, dienten mehrere Stichproben von hand- und maschinengeschriebenen Zeichen als Repräsentanten des Zufallsprozesses $\{\mathbf{v}\}$. Die Stichproben unterscheiden sich in Umfang und Schriftqualität erheblich. Sie werden im folgenden stichwortartig beschrieben:

Handblockschrift-Stichprobe Abk.: HO

Unter Beachtung von Regeln von 115 Personen geschrieben

Zeichenzahl : ca 22 000
Klassenzahl : 10
Kennungen : 1 2 3 4 5 6 7 8 9 0
Auflösung der Rasterbilder : 18 x 12

OCR-A Stichprobe Abk.: A0

Umfangreiche Stichprobe, während des praktischen Einsatzes von einer Beleglesemaschine aufgenommen; sehr breit gestreute Schriftqualität

Zeichenzahl: ca 140 000
Klassenzahl: 14
Kennungen : 0 1 2 3 4 5 6 7 8 9 H J U ±
Auflösung : 20 x 13

OCR-A Stichprobe Abk.: A1

Sorgfältig zusammengestellt breite Problemstichprobe. Schlechte Schriftqualität überrepräsentiert.

Zeichenzahl : ca 6 500
Klassenzahl : 14
Kennungen : 0 1 2 3 4 5 6 7 8 9 H J U ±
Auflösung : 20 x 13

OCR-A Stichprobe Abk.: A2

Ausgesuchte Stichprobe von schlecht gedruckten Zeichen

Zeichenzahl : ca 10 000
Klassenzahl : 10
Kennungen : 0 1 2 3 4 5 6 7 8 9
Auflösung : 20 x 13

Die im folgenden verwendeten Klassennummern beziehen sich jeweils auf die angegebene Kennungsliste.

In Bild 3 sind die der Größe nach geordneten Eigenwerte der Kovarianzmatrix **K** des Zufallsprozesses $\{\mathbf{v}\}$ als Kurve dar-

gestellt, die die Folge der N Punkte verbindet. Der Prozess $\{\mathbf{v}\}$ wird durch drei Stichproben repräsentiert.

Man erkennt einen sehr einheitlichen Verlauf aller Kurven. Bemerkenswert ist die zu Beginn (d. h., für kleine i) sehr schnelle Abnahme der Eigenwerte. So ist λ_{40} nur etwa $1/50$ von λ_1 . Da die Eigenwerte entsprechend Gleichung (39) als die jeweiligen Anteile ΔR^2 eines Merkmals zur Verminderung des Rekonstruktionsfehlers aufgefaßt werden können, wird deutlich, daß wenige Merkmale w_i überwiegend für die Rekonstruktion zuständig sind.

Die Steigung aller Kurven werden etwa ab λ_{80} bis λ_{100} sehr viel kleiner. In diesem Bereich finden wir Eigenwerte etwa gleicher Größenordnung. Ab λ_{200} fallen die Kurven wiederum sehr steil ab, ein Zeichen dafür, daß die Kovarianzmatrizen nahezu singulär sind. Möchte man den Prozess $\{\mathbf{v}\}$ bis auf ein erlaubtes, zahlenmäßig aber zu vernachlässigendes ΔR^2 vollständig rekonstruieren, so genügen hierzu etwa 210-dimensionale Merkmalsvektoren. Dieser Sachverhalt läßt sich anschaulich damit erklären, daß einige wenige Bildelemente der Rasterbilder in der gesamten Stichprobe etwa ganz schwach oder überhaupt nie besetzt sind, wie es in den Ecken des Rasterbildes zu vermuten ist, oder aber nahezu immer besetzt sind, wie etwa einige zentrale Bildelemente. Die Rangerniedrigung (- im numerischen Sinn -) kann aber auch dadurch entstehen, daß benachbarte oder beliebige Bildelemente eng miteinander korreliert sind oder sich nahezu fehlerfrei durch Linearkombination anderer Bildelemente darstellen lassen.

Um festzustellen, wieviele Merkmale zu einer hinreichenden Rekonstruktion der Zeichen notwendig sind, wurde der normierte Rekonstruktionsfehler R^2/R_0^2 in Bild 4 abgetragen über der Zahl M der verwendeten Merkmale w_i . Diese Kurven wurden für alle vier beschriebenen Stichproben bestimmt.

Dabei wurden die M Merkmale w_i , für die jeweils der Rekonstruktionsfehler R^2/R_0^2 abzulesen ist, nach der LU-Rangreihenfolge (ausgezogene Kurven) ausgewählt. Die Kurven beginnen bei dem Startwert $R^2/R_0^2 = 1$. R_0^2 ist hier definiert als der Fehler, der sich ergibt, wenn noch kein Merkmal w_i zur Rekonstruktion verwendet wird. Nach Gleichung (27) ist dies die Summe aller N Eigenwerte und entspricht der Summe aller Varianzen der Merkmale. Werden alle Merkmale zur Rekonstruktion verwendet, so ist der verbleibende Rekonstruktionsfehler 0, d. h., die Zeichen sind vollständig rekonstruierbar. Dieser Fall tritt entsprechend der Zahl der Rasterbildelemente, für die Handblockschrift bei $M = 216$ und für die OCR-A Stichproben bei $M = 260$ ein.

Es zeigt sich, daß die Werte R^2/R_0^2 der Handblockschrift-Stichprobe für kleine M etwas höher liegen als die der verschiedenen OCR-A Proben. Der grundsätzliche Verlauf aller Kurven ist jedoch sehr ähnlich. Charakteristisch ist insbesondere der sehr steile Abfall der Kurven für kleine M . So ist abzulesen, daß bei Verwendung der ersten 10 Merkmale der Rekonstruktionsfehler bereits auf 0,5 gesenkt worden ist. Bei großem M verlaufen alle Kurven sehr flach, d. h., die an dieser Stelle neu mit einbezogenen Merkmale tragen zur gesamten Bildrekonstruktion wenig bei. Ein erfahrungsgemäß tolerierbarer Fehler im Bereich von $R^2 = 25\% - 30\%$ wird für alle Stichproben bei etwa 40 Merkmalen erreicht.

Das Fehlermaß R^2 kann natürlich auch für jede andere Untermenge ausgesuchte Merkmale berechnen und nicht nur für die nach der LU-Rangordnung bestimmten. Ordnet man die Merkmale nach der MS-Rangordnung und mißt ebenfalls den jeweils zugehörigen Rekonstruktionsfehler R^2 , so erhält man die in Bild 4 unterbrochene Kurve. Sie ist die Kurve für die Stichprobe A0 und wurde stellvertretend für alle übrigen eingezeichnet.

Man sieht, daß die beiden entsprechenden Kurven sehr dicht

beieinander liegen. Offensichtlich sind beide Rangordnungen sehr ähnlich und produzieren daher nahezu die gleichen $\hat{R}^2$ -Kurven. Für alle anderen Stichproben zeigt sich das gleiche Verhalten. Damit wird deutlich, daß die Merkmale w_i , die unter dem Gesichtspunkt Klassifikation ausgewählt wurden, offenbar auch eine gute Auswahl hinsichtlich der Rekonstruktion darstellen. Dies kann anhand einiger weiterer Diagramme bestätigt werden.

Das am Rekonstruktionskriterium orientierte Fehlermaß R^2 wurde in Bild 4 sowohl nach LU wie nach MS geordneten Merkmalen bestimmt. In Bild 5 ist in entsprechender Weise das Fehlermaß S^2 - für jeweils nach beiden Rangordnungen bestimmten Untermengen von Merkmalen - über der Zahl M der Merkmale w_i abgetragen. Die Kurven wurden wiederum für die verschiedenen Stichproben gemessen. Die durchgezogenen Kurven beziehen sich dabei auf die nach der MS-Rangordnung ausgewählten Merkmale. Sämtliche Kurven beginnen bei einem von der Klassenzahl und der klassenspezifischen Auftretenswahrscheinlichkeit p_i der verschiedenen Klassen in der Stichprobe abhängigen Wert S_0^2 . Er ist

$$S_0^2 = \sum_{i=1}^K p_i(1-p_i) \quad (40a)$$

und vereinfacht sich bei gleicher Auftretenswahrscheinlichkeit $p_i = 1/K$ zu

$$S_0^2 = 1 - \frac{1}{K} \quad (40b)$$

Alle Kurven fallen zu Beginn sehr steil, knicken innerhalb eines schmalen Bereichs und verlaufen dann nahezu waagerecht. Dieser Verlauf zeigt an, daß offensichtlich einige wenige Merkmale w_i in der Lage sind, die Reststreuung S^2 erheblich zu verkleinern. Dies sind bei allen Stichproben etwa die ersten 20 ... 30 Merkmale nach der MS Rangordnung. Schon die folgenden etwa 40 Merkmale verringern die Reststreuung dann

nahezu auf den Endwert S_N^2 . Die restlichen Merkmale liefern dann nur noch verschwindend kleine Beiträge ΔS_i^2 .

Damit zeigt sich, daß die Bewertung der durch die HAT erzeugten Merkmale mit dem an der Klassifikation orientierten Maß eine deutliche Aufteilung der Menge aller Merkmale in Gruppen ermöglicht, nämlich in die Gruppe weniger, besonders wichtiger, in die Gruppe wichtiger und in die Gruppe von praktisch bedeutungslosen Merkmalen.

Für die Stichprobe A0 wurde - wiederum stellvertretend für alle Stichproben - die Kurve $S^2(M)$ für die LU-Rangordnung der Merkmale mit eingezeichnet (unterbrochene Kurve). Diese liegt sehr dicht bei der entsprechenden Reststreuungskurve mit Merkmalen im Schätzpolynom, die nach der MS-Strategie ausgesucht sind. Damit zeigt sich auch hier die sehr enge Verwandtschaft beider Auswahlstrategien.

Vergleicht man den Verlauf des Rekonstruktionsfehlers Bild 4 und der Reststreuung Bild 5, wobei beide jeweils über der Zahl der Merkmale abgetragen sind, so fällt der wesentliche Unterschied unmittelbar auf. Während die Kurven $R^2(M)$ bis zu dem Endwert $R_N^2 = 0$ stetig fallen, erreichen sämtliche $S^2(M)$ Kurven schon sehr frühzeitig ihren endgültigen Wert. Daran sieht man, daß zur Klassifikation mit dem linearen Klassifikator wenige Merkmale ausreichen. Die Steigerung der Merkmalszahl führt dann kaum noch zur Erhöhung der Erkennungsrate. Zur Rekonstruktion der Zeichen werden jedoch weit mehr Merkmale benötigt. Vollständige Rekonstruktion ist dabei unabhängig von der Stichprobe immer möglich bei der Verwendung aller Merkmale. Die erreichte Reststreuung bei Berücksichtigung aller Merkmale ist jedoch, wie das Diagramm zeigt, sehr stark von der Stichprobe abhängig. Da der Typ des Klassifikators in allen Fällen der gleiche ist, kann aus der jeweils erreichten Reststreuung S_N^2 auf die Schwierigkeit des durch die Stichprobe definierten Erkennungsproblems geschlossen werden.

S_N^2 ist für die OCR-A Stichprobe A2 am kleinsten. Danach folgt Handblockschrift, gefolgt von A1 und A0. Offenbar stellt die sehr umfangreiche Probe A0 (140 000 Zeichen), in der sowohl gute wie extrem schlecht gedruckte Ziffern vorkommen, das schwierigste Problem für den linearen Klassifikator dar. Dagegen wird die Stichprobe A1 mit relativ wenigen, aber ausgesucht schlechten Zeichen (6 500) als leichter zu klassifizieren eingeschätzt. Hier werden die Koeffizienten sowohl der Transformationsmatrix B wie der Klassifikatormatrix A_{Lin} ausschließlich auf die schlechten Zeichen adaptiert. Sicherlich werden demnach gut gedruckte Ziffern nicht mit hoher Sicherheit erkannt, da diese während der Adaptionsphase nicht vorkamen.

An die Merkmalsextraktionsstufe (1. Stufe) wurde die Forderung gestellt, daß sie Merkmale produziert, die einen minimalen Rekonstruktionsfehler R^2 garantieren. Dabei wird die Zahl $M < N$ von Merkmalen, die zur Rekonstruktion verwendet werden, jeweils vorgegeben. Nun ist es aber auch durch einfaches Weglassen von direkt gemessenen Merkmalen v_i möglich, eine Reduktion der Zahl der Merkmale zu erhalten. Die Rekonstruktion des ursprünglichen Prozesses durch die wenigen M ausgesuchten Merkmale v_j ist auch hier fehlerbehaftet. Der Rekonstruktionsfehler ergibt sich in diesem Fall daraus, wie gut der Vektor V^* der weggelassenen Merkmale durch eine lineare Schätzung

$$\tilde{V}^* = C^T \begin{pmatrix} 1 \\ V^{**} \end{pmatrix} \quad (41)$$

aus dem Vektor V^{**} der weiterverwendeten Merkmale re-produzierbar ist /01/. Der Merkmalsvektor V ist somit in zwei Vektoren V^* und V^{**} aufgeteilt worden:

$$V = \begin{pmatrix} V^{**} \\ V^* \end{pmatrix}, \quad (42)$$

Der Rekonstruktionsfehler ergibt sich nach Optimierung von C zu

$$R^2 = \min_C \left[E \left\{ |v^* - \tilde{v}^*(C)|^2 \right\} \right] \quad (43)$$

In Bild 6 ist für die Handblockschrift-Stichprobe der normierte Rekonstruktionsfehler R^2/R_0^2 in Abhängigkeit von der Zahl M der verwendeten Merkmale mit und ohne Hauptachsen-
transformation abgetragen. Daraus ist abzulesen, daß mit 40 durch die HAT erzeugten Merkmale w_i derselbe Rekonstruktionsfehler erreicht wird, wie mit 75 direkt gemessenen Merkmalen v_i .

In Bild 7 ist ebenfalls für die Stichprobe H0 das Fehlermaß S^2 über der Zahl M der verwendeten Merkmale jeweils wieder mit und ohne HAT abgetragen. Hieraus ist ersichtlich, daß für jede gewählte Anzahl M die Merkmale w_i eine kleinere Reststreuung ermöglichen.

Um die beiden diskutierten Rangordnungen (MS,LU) einander gegenüberzustellen, wurde eine Darstellung gewählt, die einen anschaulichen Eindruck von der Korrelation beider Ordnungen vermittelt (Bild 8 bis 11). Dabei stellt die Abszisse die MS-Rangnummer, die Ordinate die LU-Rangordnung dar. Jedem Merkmal kann in diesem Koordinatensystem ein Punkt zugeordnet werden. Die Koordinaten sind dabei die für dieses Merkmal ermittelte MS-LU-Rangordnungsnummer. In dieser Darstellung spiegelt sich eine enge Verwandtschaft beider Rangordnungen darin nieder, daß die zu den Merkmalen zugehörigen Punkte sich innerhalb eines schmalen Bereichs um die 45° Gerade häufen. Dies ist, zumindest für niedrige Rangnummern bei allen OCR-A Stichproben (Bild 8 bis 10) sowie für die Handblockstichprobe (Bild 11) festzustellen. Mit größer werdenden Rangnummern nimmt die Streuung der Punkte etwas zu. Am deutlichsten ist dies bei der Stichprobe A0.

Ebenso wie es möglich ist, die durch die HAT erzeugten Merkmale nach den beiden Strategien MS und LU zu ordnen, ist es auch möglich, die ursprünglichen Zeichenmerkmale v_i zu ordnen. Wählt man, um die Verwandtschaft der beiden Rangordnungen LU und MS zu demonstrieren, die gleiche Darstellung, so ergibt sich Bild 12 /01/. Die weit zerstreuten Punkte zeigen deutlich den erheblichen Unterschied der Rangordnungen an. So gibt es beispielsweise Merkmale, die unter den 10 wichtigsten bezüglich der MS-Rangordnung eingereiht wurden, während sie bezüglich der LU-Rangordnung Positionsnummern größer 150 besitzen.

Wir ziehen aus den dargestellten experimentellen Befunden folgende Schlußfolgerung: während im allgemeinen MS- und LU-Rangordnung nicht übereinstimmen (etwa bei den direkt gemessenen Merkmalen v_i), erzeugt die HAT Merkmale w_i , die gleichzeitig sowohl für die Rekonstruktion als auch für die Klassifikation nach Nützlichkeit geordnet sind. Das ist zunächst etwas erstaunlich, da diese ausschließlich unter dem Gesichtspunkt "Rekonstruktionsfähigkeit" erzeugt wurden, kann aber andererseits auch anschaulich gerechtfertigt werden.

So ist für die Klassifizierung wichtig, relevante Information zu bewahren; daher ist, was zur Rekonstruktion geeignet ist, auch geeignet für die Klassifizierung. Rekonstruktion mit nur wenigen Merkmalen w_i liefert zunächst nur eine Zeichenrekonstruktion in Umrissen /N4/. Erhöhung der Merkmalsanzahl liefert die bisher fehlende Feinstruktur nach.

Bei den üblichen Schriftzeichenformen kann aber diese Feinstruktur der Zeichen nur von untergeordneter Bedeutung sein, denn Unempfindlichkeit gegenüber Störungen der Feinstruktur ist sicherlich notwendig, damit Zeichen auch für den menschlichen Gebrauch geeignet sind. Die zur Klassifizierung wichtige Information liegt damit überwiegend in den Umrissen, d. h., der Grobstruktur der Zeichen.

2.4. Anschauliche Interpretationen zur Transformationsmatrix

Jedes durch die Hauptachsentransformation erzeugte Merkmal w_i wird aus dem Merkmalsvektor $\mathbf{V}$ durch das Skalarprodukt $\mathbf{b}_i^T (\mathbf{V} - \bar{\mathbf{V}})$ berechnet. Damit kann die Merkmalsberechnung als eine spezielle lineare Filterung des Merkmalsvektor $\mathbf{V}$ aufgefaßt werden. Das Korrelationsfilter realisiert die Funktion $\mathbf{b}_i^T \mathbf{V}$. Der entsprechende Wert w_i des Merkmals i ergibt sich schließlich durch Addition des konstanten Wertes $\mathbf{b}_i^T \bar{\mathbf{V}}$ (Bild 13).

Wie der Merkmalsvektor $\mathbf{V}$ kann nun auch der Transformationsvektor $\mathbf{b}_i$ als Rasterbild dargestellt werden, wobei allerdings die jedem Rasterbildelement zugeordneten Werte b_{ij} sowohl positive wie negative Werte annehmen können. Dieses Rasterbild kann aufgespalten werden in ein Positivbild, das nur Rasterbildelemente mit positiven Bewertungskoeffizienten besitzt, wobei die übrigen Koeffizienten auf Null gesetzt sind, und einem Negativbild, mit entsprechend negativen Werten der Bewertungskoeffizienten.

Damit wird deutlich, daß sich besonders große positive Werte für ein Merkmal w_i dann ergeben, wenn $\mathbf{V}$ an den Stellen des Rasterbildes die Komponente +1 besitzt, an denen große positive Koeffizienten von $\mathbf{b}_i$ vorliegen (gute Korrelation von $\mathbf{V}$ mit dem Positivbild). Besonders kleine Werte für w_i ergeben sich entsprechend, wenn $\mathbf{V}$ an den Stellen des Rasterbildes die Komponente +1 besitzt, an denen negative Koeffizienten von $\mathbf{b}_i$ vorliegen (gute Korrelation mit dem Negativbild). Demnach betonen die Eigenvektoren bestimmte Zeichenteile.

Welche Formelemente durch die Transformationsvektoren $\mathbf{b}_i$ hervorgehoben werden, soll die Rasterbilddarstellung verdeutlichen. In Bild 14 sind die Koeffizienten von $\mathbf{b}_i$ jeweils mit zwei Dezimalstellen (in Prozent) in die Raster-

bildelemente eingetragen. Zeichenteile ergeben sich in dieser Darstellung durch Verbinden von Rasterbildelementen mit gleichem "Potential", d. h., mit gleich großen Transformationskoeffizienten. Offensichtlich ist, daß diese Formteile durch das menschliche Auge kaum erkannt werden. Wir wählen daher aus Gründen der Anschaulichkeit eine ternär quantisierte Rasterbilddarstellung der Eigenvektoren. Dabei werden Rasterbildelemente mit positiv besetzten Werten oberhalb einer Schwelle S mit (+) gekennzeichnet, und Rasterbildelemente mit negativ besetzten Werten unterhalb $-S$ mit (-). Die restlichen Elemente werden nicht gekennzeichnet. Die Bilder 15 bis 18 zeigen diese Rasterbilddarstellung für die verschiedenen Stichproben. Die Numerierung der Eigenvektoren erfolgte dabei nach der Größe ihrer zugehörigen Eigenwerte (LU-Rangnummer der entsprechenden Merkmale w_i).

Man erkennt in den Rasterbildern mit niedriger Nummer - sie gehören zu Eigenvektoren, die wichtige Merkmale erzeugen - Formelemente von Zeichen, die für die jeweilige Schriftart typisch sind. Besonders deutlich ist dies bei Eigenvektor N^2 1 und 2 der A2 Stichprobe. Hier sind die Zeichen "7" und "J" sichtbar. Ebenso auffallend ist die "0" als Positivbild und die "1" als Negativbild in dem ersten Eigenvektor der Handblockschrift-Stichprobe H0. In diesem speziellen Fall, in dem sowohl das Positivbild wie das Negativbild ein vollständiges Zeichen darstellt, kann unmittelbar eine Aussage über die Richtung dieses Eigenvektors im N -dimensionalen Merkmalsraum gemacht werden. Dieser Vektor ist gerichtet von dem Schwerpunktsvektor der Klassenkennung "1" zu dem Schwerpunktsvektor von Klassenkennung "0."

Für alle Stichproben sind noch die Rasterbilder von Eigenvektoren hoher Nummer gezeichnet. In diesen Bildern sind nun keine Zeichenteile als Positiv- oder Negativbild zu erkennen. Der häufige Vorzeichenwechsel auch innerhalb eines kleinen Bildbereichs zeigt die hohe Ortsfrequenz an. Diese unruhigen Bilder, bei den hohen Rangnummern von Eigenvekto-

ren und die ruhigen Formen bei den Eigenvektoren mit niedriger Nummer zeigen auch die perspektivischen Darstellungen (Bild 19 bis 21, Stichprobe H0).

Mit diesen Bildern kann anschaulich das Charakteristische der Merkmalsextraktionsstufe demonstriert werden. Diese wurde so dimensioniert, daß der verbleibende Rekonstruktionsfehler R^2 minimal ist, wenn es gilt, den N-dimensionalen Zufallsprozess mit wenigen M Merkmalen zu beschreiben. Entsprechend Gleichung (3) ist das aus M Merkmalen w_i mit Fehlern rekonstruierte Bild

$$v^* = \sum_{i=1}^M w_i b_i + c$$

Bei der schrittweisen Rekonstruktion der Bilder durch Erhöhung von M werden zunächst, wie die Rasterbilddarstellung der Transformationskoeffizienten zeigt, die Formelemente von Zeichen stark betont. Damit ist eine Wiedergabe der wesentlichen Konturen der Zeichen schon mit kleiner Zahl von Summanden möglich. Erst mit zunehmender Zahl von Summanden wird Wert gelegt auf die genaue Wiedergabe der Feinstruktur der Zeichen, was durch Betonung der höheren Ortsfrequenzen erreicht wird. Damit zeigen sich deutliche Parallelen zu der zweidimensionalen Fouriertransformation. Auch hier werden die Zeichen durch Summation von "Masken" mit ständig größer werdenden Ortsfrequenzen und entsprechender Bewertung durch die Fourierkoeffizienten rekonstruiert. Während jedoch bei der Fouriertransformation die "Masken" vorgegeben sind, werden diese bei der HAT an die Aufgabenstellung, nämlich die Stichprobe als Repräsentant des Zufallsprozesses adaptiert.

3. Einblick in die Eigenschaften des zeichenerzeugenden
Prozesses $\{v, y\}$ durch Betrachtung der Merkmale w_j

Geometrisch anschauliche Deutungen spielen allgemein in der Zeichenerkennung eine äußerst wichtige Rolle. So läßt sich das Prinzip, nach dem jedes beliebige Erkennungssystem einem angebotenen Zeichen auf Grund von Meßwerten, die über dieses Zeichen vorliegen, eine mutmaßliche Klasse zuordnet, unmittelbar in die räumliche Vorstellungswelt übertragen. Hierzu denkt man sich jedem Zeichen im mehrdimensionalen Raum einen Punkt zugeordnet. Die Zuordnungsvorschrift ergibt sich unmittelbar aus den an den Zeichen vorgenommenen Messungen. Sie entspricht einer Abbildung des Zeichens, das durch eine bestimmte schwarz-weiß Verteilung auf dem Papier beschrieben ist, in diesen Raum.

Bei der Anwendung von allen Klassifizierungsmethoden interessiert nun, wie sich die Repräsentanten der Zeichen im Merkmalsraum verteilen, wie gut dieser Raum ausgenützt wird und wie demzufolge eine günstige Trennung der Klassengebiete erfolgen kann. Jede Art von Erkennungssystemen erfordert zu seiner Konstruktion Kenntnisse - in der einen oder anderen Art - über der im Merkmalsraum vorliegenden Situation. Charakteristisch ist, daß man sich dem Ziel der überhaupt erreichbaren Klassifikatorleistung umso mehr nähert, je genauere Kenntnisse vorhanden sind.

Gibt es darüberhinaus a priori Kenntnisse, d. h., müssen diese nicht erst während der Klassifikatoradaption erworben werden, so wird dadurch in der Regel eine Vereinfachung in der Berechnung des Klassifikators erreicht. Demnach interessiert insbesondere, wie gut die gemessenen Verteilungen durch möglichst einfache Modellverteilungen beschrieben werden können und wie groß der durch die Modellierung entstandene Fehler abzuschätzen ist.

Die Schwierigkeit, die einer einfachen Beschreibung der Verteilung der Punktmenge, die den zeichenerzeugenden Prozess repräsentiert, im Wege steht, ist die für praktische Anwendung notwendige hohe Zahl von Dimensionen des Raumes. In dem hier betrachteten System der Merkmalsgewinnung an Schriftzeichen werden Meßwerte durch den Abtaster gewonnen, der das gesamte Bild in Rasterbildelemente aufteilt, deren Grauwert mißt und schließlich nach einem Schwellwertvergleich binär quantisierte Merkmale abliefert. Die Zeichen werden durch Bildvorbereitungsprozeduren aufbereitet und schließlich repräsentiert durch den 200 - 250-dimensionalen binär quantisierten Merkmalsvektor $\mathbf{v}$. Wegen der binären Quantisierung der Merkmale können nur ganz bestimmte Positionen im N-dimensionalen Merkmalsraum besetzt sein. Dies sind die 2^N Ecken des N-dimensionalen Hyperkubus.

Der Mittelpunkt des Koordinatensystems $[\mathbf{v}]$ liegt in einem Eckpunkt des Kubus, die N Basisvektoren liegen auf Kanten. Der N-dimensionale Hyperkubus läßt sich durch eine Hyperkugel gleicher Dimension und mit dem Radius $R = \frac{1}{2} \sqrt{N}$ umschreiben. Sämtliche Repräsentanten der Zeichen liegen damit auf dieser Hyperkugel.

Dieses ursprüngliche Meßsystem ist zur anschaulichen Darstellung der Verteilung von Klassengebieten im Merkmalsraum ungeeignet. Die Betrachtung einer eindimensionalen Randverteilung oder einer Projektion der Punkte auf Ebenen, die durch Basisvektoren des Koordinatensystems festgelegt sind, würde keinerlei Vorstellung räumlicher Ausdehnung von Punkten einer Klasse im Merkmalsraum vermitteln, da gerade in diesen Raumrichtungen Zeichen nur bei den Werten 0 oder 1 der Skala überhaupt zu finden sind: jede Darstellung einer Randverteilung von Merkmalen müßte sich dabei beschränken auf die Angaben, wie häufig der Wert 0 bzw. 1 als Zeichenmerkmal vorkommt (Zweipunkteverteilung). Die Beziehung dieser Merkmale untereinander, durch die charakteristische Formen von Zeichen zustande kommen, bleiben vollständig unbe-

rücksichtigt. Ein Merkmal - völlig isoliert von den übrigen 200 betrachtet - ermöglicht damit sicherlich keine Aussagen über charakteristische Eigenschaften des Prozesses. Dies ist erst möglich, wenn gegenseitige Bindungen einer größeren Menge von Merkmalen mit berücksichtigt werden.

Nun erweist sich das durch die HAT gegebene Koordinatensystem als besonders geeignet für Interpretationen zum Prozess $\{\mathbf{v}, \mathbf{y}\}$, da es an diesen Prozess angepaßt ist (problemorientiertes, maßgeschneidertes Koordinatensystem). Daher verwenden wir hier das in Abschnitt 2.1. eingeführte System und betrachten den Resultatsprozess der Merkmale $\{\mathbf{w}\}$, um daraus Rückschlüsse auf den Prozess $\{\mathbf{v}\}$ zu ziehen.

Die Darstellung von Randverteilungen ebenso wie von Projektionen der Punkte auf Ebenen, die von den Basisvektoren des neuen Systems aufgespannt sind, kann gedeutet werden als die Darstellung des Prozesses aus einer neuen Perspektive. Durch günstig gewählte Perspektiven wird versucht, die räumliche Verteilung besser auszuleuchten. Wie wichtig die Wahl von Standort und Blickrichtung ist, wird unmittelbar am Beispiel der zweidimensionalen Darstellung von dreidimensionalen Objekten durch die Photographie deutlich. Eine geeignet gewählte Perspektive kann hier einen guten räumlichen Eindruck vermitteln.

Die Koordinaten der Punkte in dem neuen Meßsystem sind die Merkmale w_i . Sie ergeben sich unmittelbar als Linearkombinationen der ursprünglichen Meßwerte v_i . Damit wird bei der Beschreibung des Zufallsprozesses durch ein Merkmal berücksichtigt, daß erst mehrere Meßwerte v_i in der Lage sind, charakteristische Zeichenteile zu bilden. Die klassenweise Verteilung eines einzigen Merkmals w_i ist daher schon aussagefähig.

Die Basisvektoren des HAT-Systems orientieren sich an den

Streuungen der Punktmenge in den verschiedenen Raumrichtungen. Die Randverteilung des Prozesses in Richtung des i 'ten Basisvektors entspricht der Wahrscheinlichkeitsdichteverteilung des Merkmals w_i . Werden die Zeichen nur durch die Merkmale w_i und w_j beschrieben, und diese Werte als Ordinate und Abszisse von Punkten in ein Koordinatensystem eingetragen, so entspricht dies einer Projektion des N -dimensionalen Prozesses auf eine Ebene, die von dem Basisvektor i und j aufgespannt wird.

Wird die N -dimensionale Hyperkugel, auf der die Punkte als Repräsentanten aller überhaupt existierenden Zeichen zu liegen kommen müssen, projiziert auf eine Ebene, so erhält man die von einem Kreis umschlossene Fläche. Sämtliche Punkte bilden sich somit auf diese Fläche ab. Dabei sind "nur" die $2^N = 2^{200} \approx 1,6 \cdot 10^{60}$ Positionen als Bilder der Ecken des N -dimensionalen Hyperkubus theoretisch besetzbar. Auf welche Stelle jeder Punkt projiziert wird, ist allein von der gewählten Projektionsebene, d. h., von den Basisvektoren des neuen Koordinatensystems abhängig.

Da allein schon die Zahl der in einer Stichprobe enthaltenen Elemente wesentlich kleiner als die Zahl 2^N besetzbarer Positionen ist, müssen nicht durch Zeichen besetzte Gebiete vorhanden sein. Hinzu kommt noch, daß sehr viele im Prinzip mögliche Rasterbilder nicht als sinnvolle Zeichenbilder vorkommen. Es interessiert somit, wie realistische Zeichen abgebildet werden, wie gedrängt Zeichen einer Klasse liegen, und wie gut der Merkmalsraum ausgenutzt wird.

Hierzu wird der Merkmalsraum auf die vom ersten und zweiten Eigenvektor aufgespannte Fläche projiziert. Dies ist die größte und zweitgrößte Ausdehnung der Punktmenge. Für das Beispiel der Handblockschriftstichprobe H0 ergibt sich dabei Bild 22. Von den 10 Klassen sind nur Zeichen von denjenigen Klassen eingezeichnet, die die betragsgrößten Werte der Merkmale w_i besitzen. Dies sind die Zeichen mit

der Kennung "0", "1", "2". Sie bilden die äußere Umrandung aller Zeichen der Stichprobe. Aus dieser Darstellung wird deutlich, daß der Merkmalsraum bereits in den Richtungen größter Ausdehnung der Punkte schlecht ausgenutzt ist. Sämtliche Punkte konzentrieren sich im wesentlichen innerhalb eines Kreises, dessen Radius $\frac{2}{3} R$ beträgt. Der Mittelpunkt des Hyperkubus ($\mathbf{V} = (0.5, 0.5, \dots, 0.5)$) kommt etwas außerhalb des Schwerpunktes der Punktmenge, der dem Koordinatenursprung entspricht, zu liegen. Mit eingezeichnet ist der Punkt, in den der Mittelpunkt des ursprünglichen Systems ($\mathbf{V} = (0, 0, \dots, 0, 0, 0)$) abgebildet wird. Dieser Punkt liegt nahe bei den Zeichen der Kennung "1". In dieser Darstellung fällt die relativ scharfe, etwa kreisbogenförmige Kontur auf, die das Klassengebiet nach außen abgrenzt. Während noch sehr viele Zeichen aller Klassen in Richtung Koordinatenursprung weitab von dem Klassenschwerpunkt zu finden sind, ist das äußere Gebiet vollkommen leer. Wird der Mittelpunkt des Kreises als Nord- bzw. als Südpol der Hyperkugel aufgefaßt, so kann anschaulich die Zone, in der Zeichen zu finden sind, definiert werden. Es ist dies eine ausgedehnte Polkappe auf der nördlichen Halbkugel, die durch einen Breitenkreis nach Süden begrenzt ist. Auf der entsprechenden südlichen Polkappe würden die Negativbilder realer Zeichen zu liegen kommen.

3.1. Eindimensionale Randverteilung

Im folgenden werden Messungen an eindimensionalen Randverteilungen des Zufallsprozesses in den verschiedenen Raumrichtungen, die durch die Eigenvektoren definiert sind, ausgeführt. Diese Randverteilungen sind die Wahrscheinlichkeitsdichtefunktionen der Merkmale w_i . Sie wurden ausgemessen für die Handblockschriftstichprobe H0 sowie für die OCR-A Stichprobe A0. Die Darstellung erfolgt jeweils als Histogramm, wobei die Ordinate die absolute Anzahl der gemessenen Ereignisse angibt.

Die Bilder 23 bis 25 zeigen die gemessenen klassenunabhängigen Wahrscheinlichkeitsdichtefunktionen

$$p(w_i) \quad i = 1, 2, 3$$

für die Handblockschriftstichprobe H0. Da das Koordinatensystem im Schwerpunkt liegt, ist der Mittelwert jeweils 0. Die Varianz ist der zu dem jeweiligen Merkmal zugehörige Eigenwert λ_i . An den gemessenen Dichtefunktionen fällt der teilweise recht abrupte Abbruch auf. Dieser steile Abfall der Dichte nach außen wurde in Bild 22 bereits festgestellt. Besonders ausgeprägt sind bei der Dichtefunktion des 1. Merkmals 2 Maxima. Offensichtlich liegen an diesen Orten im Merkmalsraum die Dichtezentren mehrerer Klassengebiete. Die mit eingezeichneten Normalverteilungen mit gleichem Mittelwert und gleicher Varianz demonstrieren die Abweichungen von der Symmetrie.

In Bild 26 sind die Hüllkurven der Dichtefunktionen von Merkmal 1, 10, 20, 30, 40 in ein Diagramm gezeichnet. Hier wird die unterschiedliche Streuung in den verschiedenen Richtungen des Merkmalsraumes deutlich. Da die Integration entlang den Koordinatenachsen w_i für alle Kurven denselben Wert ergibt - bei der hier verwendeten nicht normierten Darstellung ist die Summe der Ordinaten der Meßpunkte die Anzahl der Stichproben - müssen die Dichtefunktionen mit kleinerer

Streuung deutlich größere Werte annehmen. Aus dem Diagramm kann abgelesen werden, daß die Dichte um den Ursprung des Koordinatensystems in der Richtung des 1. Basisvektors nur etwa $\frac{1}{5}$ der Dichte in Richtung des 40. Basisvektors beträgt. Bezeichnend für die schlanken Dichtefunktionen der Merkmale höherer Rangordnung ist die Symmetrie um den Ursprung. Bild 27 zeigt die perspektivische Darstellung der Verteilung der ersten 30 Merkmale.

Bisher haben wir die Randverteilung ohne Aufschlüsselung nach den verschiedenen Klassen betrachtet. Mit Aufschlüsselung nach den Klassen ergeben sich die klassenspezifischen Randverteilungen $p(w_i/k)$, die in den Bildern 28 bis 32 dargestellt sind. Diese geben typische Formen aller ausgemessenen Verteilungen wieder und können somit stellvertretend für alle Dichtefunktionen stehen. In jedes Diagramm wurde als Modellkurve eine Normalverteilung (NV) mit gleichem Mittelwert und gleicher Standardabweichung eingezeichnet. Wir stellen fest, daß es klassenspezifische Randverteilungen gibt, die sich gut, und andere, die sich weniger gut durch Normalverteilungen approximieren lassen. Zu denen, die große Ähnlichkeiten mit der NV aufweisen, gehören die Verteilungen auf den Bildern

29 (Dichteverteilung w_1 , Klasse 2)

31 (Dichteverteilung w_3 , Klasse 2)

32 (Dichteverteilung w_{13} , Klasse 1)

Größere Abweichungen registrieren wir dagegen bei den Diagrammen

28 (Dichteverteilung w_1 , Klasse 1)

30 (Dichteverteilung w_3 , Klasse 1)

wobei die Abweichung in Bild 28 besonders groß ist. In diesem Beispiel ergibt die Messung der maximalen Dichte einen gegenüber der NV um 30% höheren Wert.

Die gemessene Dichteverteilung des Merkmals 3 (Kennung "1") weist ihr Maximum deutlich versetzt vom Mittelwert auf. Auch diese Verteilung eignet sich daher wenig für eine Modellierung mit der NV. Der prinzipielle Verlauf aller gemessenen Dichteverteilungen ist jedoch sehr ähnlich: sämtliche Kurven sind hier bei der Stichprobe H0 unimodal, häufig einigermaßen symmetrisch, weiche Formen überwiegen.

Gleichartige Messungen, wie die eben diskutierten, sind für die OCR-A Stichprobe A0 durchgeführt worden, die mit 10 000 Zeichen pro Klasse außerordentlich umfangreich ist. Gemessen wurde dabei die

Dichteverteilung w_1	, Klasse 1	(Bild 33)
Dichteverteilung w_1	, Klasse 3	(Bild 34)
Dichteverteilung w_1	, Klasse 4	(Bild 35)
Dichteverteilung w_2	, Klasse 3	(Bild 36)
Dichteverteilung w_{10}	, Klasse 3	(Bild 37)
Dichteverteilung w_{10}	, Klasse 4	(Bild 38)

Nach ihrem Aussehen lassen sich diese Verteilungen in zwei Gruppen einteilen, nämlich in die Gruppe der unimodalen Verteilungen - dies sind die Verteilungen des Merkmals w_{10} -, und in die Gruppe nicht unimodaler Kurven, in der die restlichen Verteilungen einzuordnen sind. Der unruhige Verlauf der klassenspezifischen Dichteverteilung, die neben dem Hauptmaximum bis zu 3 lokale Nebenmaxima aufweist, tritt nur bei der Verteilung von Merkmalen mit großer Streuung auf, d. h., bei Merkmalen, die am Anfang der Rangordnungsliste stehen. Er kann damit begründet werden, daß neben den gut gedruckten OCR-A Zeichen solche zu finden sind, die bestimmte, teilweise stark ausgeprägte Verschmutzungs- bzw. Verstümmelungsarten aufweisen. Diese Zeichenverstümmelungen lassen sich unmittelbar mit einem schlechten Wartungszustand des Druckwerkzeugs begründen. Zeichen mit diesen typischen Verstümmelungen häufen sich dann bei bestimmten

Werten der Merkmale w_{i1} . Diese Verteilungen lassen sich nicht mehr angemessen durch die Normalverteilung (die unimodal ist) approximieren. Dagegen erweist sich die NV als geeignetes Modell für Merkmale kleinerer Streuung (z. B. w_{10}).

Vergleicht man die von der Handblockschrift erzeugten klassenspezifischen Dichteverteilungen mit denjenigen der OCR-A Stichprobe, so kann man feststellen, daß sich die Randverteilung der Merkmale bei der Handblockschrift eher für eine Modellierung durch die NV eignet, als bei der OCR-A Stichprobe.

Diese Aussagen hinsichtlich der Modellierbarkeit der einzelnen eindimensionalen Randverteilungen der klassenspezifischen Wahrscheinlichkeitsdichtefunktionen basieren ausschließlich auf der Betrachtung der Bilder, in denen gemessene Verteilung und Modellverteilung (Normalverteilung) dargestellt sind. Um diese Thesen zu erhärten, wurde mit den gemessenen Verteilungen ein statistischer Test, der Kolmogoroff-Smirnow-Test (KS-Test) durchgeführt. Dieser wird in /K5/ beschrieben und wurde bereits in /N1/ für Merkmale $w_{i/k}$ von handgeschriebenen Ziffern mit etwa 90 Zeichen pro Klasse durchgeführt. Mit diesem Test kann geprüft werden, ob Zufallsvariable, und diese sind hier die Merkmale $w_{i/k}$, aus einer Grundgesamtheit mit einer bestimmten Verteilung entstammen oder nicht. Wird angenommen, die Verteilungsfunktion des Modellprozesses sei $F(x)$, wobei x die Zufallsvariable ist, und die Verteilungsfunktion, die durch die Stichprobe ermittelt worden ist, sei $\tilde{F}(x)$, so wird beim KS-Test nach der betragsmäßig größten Abweichung

$$A = \max_x \left| F(x) - \tilde{F}(x) \right|$$

gesucht. Es ist nun eine Signifikanzzahl α vorzugeben, die angibt, mit welcher Wahrscheinlichkeit bei angenommener Gültigkeit der Hypothese ein Wert A auftreten kann, der

größer als eine Schranke A' ist. Die zu verschiedenen α zugehörigen Schranken A' , die von dem Stichprobenumfang abhängig sind, liegen tabelliert in /K5/ vor. Wählt man α klein, so trifft das Ereignis $A > A'$ selten, bei großem α entsprechend häufiger ein. Die Hypothese, die Zufallsvariable x entstamme der Grundgesamtheit mit der Verteilungsfunktion F wird angenommen, wenn der Wert A kleiner ist als die Schwelle A' , sie wird zurückgewiesen, wenn A größer A' ist.

Dieser Test wurde durchgeführt für die Merkmale $w_{i/k}$ der Handblockschrift-Stichprobe HO ($i = 1 \dots 40$, $k = 1 \dots 10$) und für die OCR-A Stichprobe AO ($i = 1 \dots 10$, $k = 1 \dots 6$). Modellprozess für $\{w_{i/k}\}$ ist wiederum die Normalverteilung. Die Ergebnisse sind in Bild 39 dargestellt.

Für jedes Merkmal w_i wurden die k klassenspezifischen Prozesse $\{w_{i/k}\}$ getestet. Es ergeben sich damit für jedes Merkmal k Abweichungen A . Wir bezeichnen die größte dieser k Abweichungen mit A^* , die kleinste mit A^{**} . Diese beiden Werte wurden jeweils in die Diagramme eingezeichnet und durch einen vertikalen Balken verbunden. Die von der Stichprobenanzahl abhängige Schwelle A' wurde jeweils für $\alpha = 1\%$ und 20% eingetragen. Die Schwellen liegen bei der Stichprobe HO wegen der geringen Probenzahl entsprechend höher.

Wie aus dem unteren Diagramm von Bild 39 erkennbar ist, liegen sämtliche Werte A^* deutlich oberhalb derjenigen Schwelle A' , die zu $\alpha = 1\%$ gehört. Die kleinsten Werte A^{**} liegen mit Ausnahme von zwei Abweichungen ebenfalls oberhalb der Schwelle A' für $\alpha = 1\%$. Damit muß die These, die Merkmale $w_{i/k}$ der OCR-A Stichprobe AO entstammen einem normalverteilten Prozess, verworfen werden.

Bei der Handblockschriftstichprobe HO (oberes Diagramm) liegen nur die Werte A^* für kleine Indizes i deutlich oberhalb der Schwelle A' , sämtliche Werte A^{**} kommen jedoch unterhalb

von A' zu liegen. Damit kann die These, $\{w_{i/k}\}$ sei normalverteilt, für einige klassenspezifische Verteilungsdichten angenommen werden, für einige muß sie jedoch verworfen werden. Aus dem Diagramm ist weiter erkennbar, daß für große Indizes i eine wesentlich bessere Übereinstimmung von Modellprozess und gemessenem Prozess gegeben ist. Sämtliche Schlußfolgerungen, die bisher hinsichtlich der Modellierbarkeit der Prozesse $\{w_{i/k}\}$ gezogen wurden, werden damit durch den KS-Test bestätigt.

3.2. Zweidimensionale Projektionen

Die Betrachtung eindimensionaler Projektionen (Randverteilungen des Prozesses) von vieldimensionalen Verteilungen ermöglicht nur einen sehr unvollkommenen Einblick in die im Merkmalsraum vorliegende Situation. Die unmittelbare Anschauung wird dagegen stärker angesprochen, wenn man die Projektionen auf zweidimensionale Unterräume durchführt.

Wir verwenden dazu Projektionsebenen, die durch jeweils zwei verschiedene Basisvektoren b_i, b_j des HAT-Koordinatensystems aufgespannt werden. Man erhält diese Darstellung, indem man zusammengehörige Werte w_i, w_j in ein rechtwinkliges Koordinatensystem einträgt. Die Dichte der Punkte wird dabei durch die Zahl der lokal vorhandenen Punkte pro Flächeneinheit wiedergegeben.

Die Bilder 40 bis 43 geben die Ebene w_1, w_2 wieder, auf die die Punkte der Kennung "0", "1", "2", "6" projiziert wurden. In allen Bildern erkennt man ein Dichtezentrum, das bei den Zeichen der Kennung "1" besonders deutlich ausgeprägt ist. Bei allen Verteilungen ist eine starke Abnahme der Dichte nach außen, d. h., zu größeren $|w_i|, |w_j|$ Werten, festzustellen. Dies bedeutet, daß das Gebiet einer Klasse nach außen schärfer abgegrenzt ist als nach innen, da in Richtung Koordinatenursprung einige verstreut liegende Zeichen der Klasse weitab vom Klassenschwerpunkt zu fin-

den sind. Damit drängen sich die gedachten Linien gleicher Wahrscheinlichkeitsdichte bei allen Bildern relativ stark nach außen, während sie in Richtung Koordinatenursprung deutlicher auseinanderliegen. Wegen der kometenartigen Schweife, die in Bild 40 zu erkennen sind, haben die Linien gleicher Wahrscheinlichkeitsdichte in diesem Bild die Form von "Hörnchen".

Bild 44 stellt die Projektion der Zeichen mit der Kennung "7" auf die Ebene $w_{10}w_{20}$ dar. Im Gegensatz zu den Bildern 40 bis 43 ist hier im wesentlichen eine gleichmäßigere Verteilung erkennbar. Die Dichten nehmen vom Zentrum nach außen etwa gleich schnell ab. Die geometrischen Orte gleicher Wahrscheinlichkeitsdichte lassen sich eher durch Ellipsen annähern als in den Bildern 40 bis 43. Zu beachten ist hier noch der geänderte Darstellungsmaßstab. Das Bild vermittelt den Eindruck, daß sich diese Verteilung einigermaßen gut durch die Normalverteilung approximieren läßt.

Um einen Überblick über die gegenseitige Lage der Klassengebiete aller Zeichenklassen in einigen, durch die Basisvektoren definierten Ebenen zu bekommen, ist es zweckmäßig, die Klassengebiete durch Modellkurven zu beschreiben. Angesichts der unregelmäßigen Formen dieser Verteilungen ist dies natürlich nur mit Vergrößerung möglich.

Wir modellieren die hier beobachteten Verteilungen durch unimodale schiefelliptische Verteilungen, bei denen Konturen konstanter Wahrscheinlichkeitsdichte Ellipsen sind, und deren Wahrscheinlichkeitsdichte mit zunehmendem Abstand vom Dichtezentrum, dem Klassenschwerpunkt $\bar{w}_k$, nach einem beliebigen monotonen Abklinggesetz abfällt.

$$p(w/k) \sim f(Q_k) \quad (44)$$

mit

$$Q_k = (\mathbf{w} - \bar{\mathbf{w}}_k)^T \mathbf{K}_k^{-1} (\mathbf{w} - \bar{\mathbf{w}}_k) \quad (45)$$

Das Dichtezentrum liegt bei

$$\bar{\mathbf{w}}_k = E\{\mathbf{w}/k\} \quad (46a)$$

Die Kernmatrix der quadratischen Form ist die Inverse der klassenspezifischen Kovarianzmatrix

$$\mathbf{K}_k = E\{(\mathbf{w} - \bar{\mathbf{w}}_k)(\mathbf{w} - \bar{\mathbf{w}}_k)^T / k\} \quad (46b)$$

Als Sonderfall ist in dieser Familie von Verteilungen, die von einer quadratischen Form abhängen, die Normalverteilung enthalten. Dann ist

$$f(Q_k) = e^{-\frac{1}{2}Q_k} \quad (47)$$

Der Wert der klassenspezifischen quadratischen Form $Q_k(\mathbf{w})$ kann als ein spezielles (schieffelliptisches) Abstandsmaß, bezogen auf den Schwerpunkt der betreffenden Klasse, aufgefaßt werden. Man bezeichnet dieses spezielle Abstandsmaß auch als Mahalanobis-Abstand.

Für $Q_k = \text{const.}$ ergibt sich - in M Dimensionen - ein Ellipsoid mit dem Mittelpunkt $\bar{\mathbf{w}}_k$. Da der Erwartungswert

$$\begin{aligned} \bar{Q}_k &= E\{(\mathbf{w} - \bar{\mathbf{w}}_k)^T \mathbf{K}_k^{-1} (\mathbf{w} - \bar{\mathbf{w}}_k) / k\} \\ &= E\{ \text{Spur} [(\mathbf{w} - \bar{\mathbf{w}}_k)(\mathbf{w} - \bar{\mathbf{w}}_k)^T \mathbf{K}_k^{-1} / k] \} \\ &= \text{Spur} \left[E\{(\mathbf{w} - \bar{\mathbf{w}}_k)(\mathbf{w} - \bar{\mathbf{w}}_k)^T / k\} \cdot \mathbf{K}_k^{-1} \right] \quad (48) \end{aligned}$$

$$\bar{Q}_k = \text{Spur} [\mathbf{K}_k \cdot \mathbf{K}_k^{-1}] = \text{Spur} [\mathbf{I}] = M$$

hier den Wert $M = \text{Dim} \{\mathbf{w}\}$ annimmt, ist es zweckmäßig,

die Ellipsoide mit dem Wert $Q_k = M$ als Repräsentanten der Klasse k zu verwenden.

Der Mittelpunkt des Ellipsoides nach Gleichung (45) ist der Schwerpunkt $\bar{w}_k$ der Klasse k . Die Richtungen der Halbachsen des Ellipsoides sind dabei durch die Eigenvektoren von K_k^{-1} gegeben. Diese sind identisch mit den Eigenvektoren von K_k . Die Halbachsen ergeben sich unmittelbar aus den Eigenwerten von K_k und lassen sich wie folgt bestimmen.

Es sei B_k die Matrix der Eigenvektoren von K_k . Dann gilt

$$B_k^T K_k B_k = L_k \quad (49a)$$

$$K_k = B_k L_k B_k^T \quad (49b)$$

$$K_k^{-1} = B_k L_k^{-1} B_k^T \quad (49c)$$

$$L_k = \begin{pmatrix} \lambda_{1k} & & \\ & \ddots & \\ & & \lambda_{mk} \end{pmatrix} \quad (49d)$$

$\lambda_{ik} = i$ 'ter Eigenwert von K_k

Gleichung (49c) in Gleichung (45) eingesetzt ergibt

$$Q_k = (w - \bar{w}_k)^T B_k L_k^{-1} B_k^T (w - \bar{w}_k) \quad (50)$$

Wir definieren eine neue Zufallsvariable

$$z = (z_1, z_2, \dots, z_M)^T \quad (51)$$

durch die Gleichung

$$z = B_k^T (w - \bar{w}_k) \quad (52)$$

Damit erhalten wir die Gleichung des M-dimensionalen Ellipsoides in Normalform

$$Q_k = \mathbf{z}^T L_k^{-1} \mathbf{z} \quad (53a)$$

$$Q_k = \frac{z_1^2}{\lambda_{1k}} + \frac{z_2^2}{\lambda_{2k}} + \dots + \frac{z_M^2}{\lambda_{Mk}} \quad (53b)$$

Somit ergibt sich als Wert für die i'te Halbachse a_{ik} des Ellipsoides allgemein

$$a_{ik} = \sqrt{\lambda_{ik} \cdot Q_k} \quad (54a)$$

und speziell für das Ellipsoid, das ein Klassengebiet beschreiben soll ($Q_k = \overline{Q_k} = M$)

$$a_{ik} = \sqrt{\lambda_{ik} \cdot M} \quad (54b)$$

Wir bestimmen die klassenspezifischen äquivalenten Ellipsen im zweidimensionalen Raum, indem wir für jede Klasse in der gewählten Darstellungsebene

$$\mathbf{w} = \begin{pmatrix} w_i \\ w_j \end{pmatrix}$$

Erwartungswert $\overline{\mathbf{w}}_k$ und Kovarianzmatrix K_k berechnen. Die Ellipsen sind mit in die Bilder 40 bis 44 eingezeichnet. Man erkennt aus allen Bildern, daß die Mehrzahl aller Punkte innerhalb der Ellipsen zu finden ist. Allerdings ist erkennbar, daß besonders bei den nichtelliptischen Verteilungen diejenigen Zeichen, die weitab von Klassenschwerpunkt liegen, nicht mit erfaßt werden. Die Wertevergrößerung des Parameters Q_k der Modellellipse würde zwar einerseits mehr Zeichen mit einbeziehen, andererseits aber den Ort der größten Konzentration zu ungenau wiedergeben.

In den Bildern 45 bis 48 sind die 10 Ellipsen als Repräsentanten der 10 Klassengebiete für das Handblockschrift-Beispiel gezeichnet, und zwar in den Ebenen

$$w_1 w_2, w_1 w_3, w_1 w_{10}, w_1 w_{30}.$$

Jede Ebene ist dabei durch die zugehörigen orthogonalen Basisvektoren definiert, nämlich den in allen Darstellungen enthaltenen Basisvektor b_1 , der die Richtung der größten Ausdehnung der Punktmenge anzeigt, und den Basisvektoren b_2 , b_3 , b_{10} , b_{30} . Damit können diese Darstellungen als eine Art Grund- und Aufriß interpretiert werden. Sie unterstützen dadurch die räumliche Vorstellung von mehrdimensionalen Klassengebieten im Merkmalsraum.

Aus allen Bildern erkennt man, daß sich als Repräsentanten der Klassengebiete verschieden orientierte und verschieden große Ellipsen ergeben, sowohl hinsichtlich der Fläche wie des Halbachsenverhältnisses. Von Bild 48 abgesehen ergeben sich auch sehr unterschiedliche Richtungen der Halbachsen. Dies weist unmittelbar auf unterschiedliche klassenspezifische Kernmatrizen der quadratischen Form in diesen Ebenen. Ebenfalls deutlich zu erkennen ist die verschiedene Ausdehnung der Klassengebiete in die jeweiligen Raumrichtungen. So ist die $w_1 w_2$ Ebene mit Klassengebieten gut ausgefüllt, dagegen reihen sich in der $w_1 w_{30}$ Ebene die Ellipsen innerhalb eines schmalen Bandes entlang der w_1 Achse. Im Gegensatz zu den übrigen Bildern sind hier die Halbachsen nahezu parallel zu den Koordinaten ausgerichtet. Bei allen Ellipsen ist eine Halbachse dominierend.

Die Darstellungen vermitteln den Eindruck einer starken Überlappung der Klassengebiete. Offensichtlich ist in zwei Dimensionen eine Trennung aller Klassengebiete voneinander kaum möglich. Ebenen, die die Klassengebiete besser voneinander trennen, können danach nur in Merkmalsräumen von höherer Dimension erwartet werden.

Aus den Darstellungen erkennt man, daß die Richtung b_1 diejenige ist, in der Zeichen mit der Kennung "0" und "1" sehr weit auseinander liegen. Der Differenzvektor ihrer Klassenschwerpunktsvektoren liegt in etwa in Richtung b_1 . Dies bestätigt die Schlußfolgerung, die aus der Interpre-

tation der Rasterbilddarstellung des ersten Eigenvektors (Bild 18) gezogen wurden. In diesem Rasterbild ist deutlich das Zeichen "1" und "0" zu erkennen, und zwar eine "1" mit negativen Komponenten (Senke), eine "0" mit positiven Komponenten (Erhebung).

Aus Bild 45 ist ersichtlich, daß das Klassengebiet der "2" in einem Randgebiet liegt, und dabei kaum mit anderen Klassengebieten in Berührung kommt. Alle Punkte in diesem Gebiet besitzen stark negative Komponenten w_2 . Auch dies ist aus der Rasterbilddarstellung des 2. Eigenvektors zu sehen. Deutlich erkennt man das Zeichen "2" als Senke, entsprechend den negativen Komponenten von w_2 .

3.3. Vergleich mit der M-dimensionalen Normalverteilung

Bei der Betrachtung der eindimensionalen Randverteilungen hat sich in zahlreichen Fällen die Normalverteilung als Modell angeboten. Wir haben dabei festgestellt, daß - wenn auch nicht in allen, so doch in den meisten eindimensionalen Projektionen - die unmittelbare Anschauung und der KS-Test eine gute Übereinstimmung zwischen dem vorliegenden Histogramm und der Normalverteilung erkennen lassen. Ebenfalls konnte festgestellt werden, daß manche zweidimensionale Projektionen eine Ähnlichkeit mit der NV aufweisen.

Die eigentlich interessierende Fragestellung ist nun allerdings nicht, inwieweit ein- und zweidimensionale Randverteilungen des Zufallsprozesses sich durch die NV modellieren lassen, vielmehr ist zu klären, ob die M-dimensionale NV geeignet ist, den Prozess in M Dimensionen gut zu beschreiben. Zwar gilt, daß bei einer M-dimensionalen Normalverteilung sämtliche eindimensionalen Randverteilungen ihrerseits notwendig wieder normalverteilt sind. Jedoch genügt die Betrachtung einiger Randverteilungen alleine nicht, um eine Aussage darüber machen zu können, ob eine

M-dimensionale Verteilung "normal" ist oder nicht.

Um dies zu prüfen, betrachten wir die Verteilung der Meßpunkte in Ellipsenringen. Die innere Ellipse soll dabei entsprechend Gleichung (45) durch den Parameter Q_k , die äußere durch $Q_k + \Delta Q_k$ definiert sein, wobei $\Delta Q_k \rightarrow 0$ gehen soll. Die Verteilung in den Ellipsenringen ist damit die Wahrscheinlichkeitsdichteverteilung des Abstandsmaßes Q_k .

Um wiederum eine anschauliche Gegenüberstellung hinsichtlich der Übereinstimmung von gemessener und modellierter Verteilungsdichte zu erhalten, stellen wir die Verteilung des nach Gleichung (45) definierten Abstandsmaßes als Histogramm dar. Wird nun angenommen der Prozess $\{w/k\}$ sei in M Dimensionen normalverteilt

$$p(w/k) = NV(\bar{w}_k, K_k)$$

so ist das verallgemeinerte Abstandsmaß Q_k , wie im folgenden gezeigt wird, χ^2 verteilt: $p\{Q_k\} = \chi^2$ verteilt, und kann somit unmittelbar in das Diagramm zum Vergleich mit eingezeichnet werden.

$$\text{Sei } p(w/k) = \frac{1}{\sqrt{2\pi}^M |K_k|} e^{-\frac{1}{2}(w-\bar{w}_k)^T K_k^{-1} (w-\bar{w}_k)} \quad (55)$$

$$\begin{aligned} &= NV(\bar{w}_k, K_k) \\ \text{dann ist } p(Q_k) &= \begin{cases} \frac{1}{2^{M/2} \Gamma(\frac{M}{2})} Q_k^{\frac{M}{2}-1} e^{-\frac{Q_k}{2}}, & Q_k > 0 \\ 0 & \text{für } Q_k \leq 0 \end{cases} \end{aligned} \quad (56)$$

$$= \chi^2 \text{ verteilt}$$

$$\text{wobei } q_k = (\mathbf{w} - \bar{\mathbf{w}}_k)^T K_k^{-1} (\mathbf{w} - \bar{\mathbf{w}}_k)$$

(siehe Gleichung (45))

Im folgenden wird bei allen Gleichungen teilweise auf den Klassenindex k verzichtet, um eine übersichtlichere Darstellung zu erhalten. Eine Verwechslung ist nicht möglich, da nur Merkmalsvektoren der Klasse k zugelassen werden. Wir führen eine Hilfsvariable $\mathbf{u}$ ein, die sich durch eine lineare Abbildung aus der Variablen $\mathbf{w}$ ergibt.

$$\mathbf{u} = \mathbf{C}^T (\mathbf{w} - \bar{\mathbf{w}}_k) \quad (57)$$

wobei gelten soll

$$\mathbf{C}\mathbf{C}^T = K_k^{-1} \quad (58)$$

Diese Zerlegung einer Matrix in ein dyadisches Produkt ist nicht eindeutig. So existieren beliebig viele Matrizen $\mathbf{C}$, die diese Bedingung erfüllen. Als Beispiel sei hier nur die "Dreieckszerlegung" (- $\mathbf{C}$ besitzt Dreiecksform -) und die "Spektralzerlegung" (- $\mathbf{C}$ wird aus den Eigenvektoren und Eigenwerten von K_k^{-1} bestimmt -) erwähnt /Z1/. Da $\mathbf{w}$ in M Dimensionen normalverteilt ist, muß auch $\mathbf{u}$ in M Dimensionen normalverteilt sein /M2, R1/. Die Normalverteilung von $\mathbf{u}$ ist vollständig beschrieben durch deren Mittelpunkt und Kovarianzmatrix:

$$\bar{\mathbf{u}} = E\{\mathbf{u}\} = \mathbf{C}^T E\{\mathbf{w} - \bar{\mathbf{w}}_k\} \quad (59)$$

$$= \mathbf{C}^T [E\{\mathbf{w}\} - \bar{\mathbf{w}}_k] = \mathbf{0}$$

$$\text{KOV}\{\mathbf{u}\} = E\{\mathbf{u}\mathbf{u}^T\} = E\{\mathbf{C}^T (\mathbf{w} - \bar{\mathbf{w}}_k) (\mathbf{w} - \bar{\mathbf{w}}_k)^T \mathbf{C}\}$$

$$\text{KOV}\{\mathbf{u}\} = \mathbf{C}^T \underbrace{E\{(\mathbf{w} - \bar{\mathbf{w}}_k)(\mathbf{w} - \bar{\mathbf{w}}_k)^T\}}_{\mathbf{K}_k} \mathbf{C}$$

$$\text{KOV}\{\mathbf{u}\} = \mathbf{C}^T \mathbf{K}_k \mathbf{C} \quad (60)$$

$$\begin{aligned} &\stackrel{\text{Gl. 58}}{=} \mathbf{C}^T (\mathbf{C} \mathbf{C}^T)^{-1} \mathbf{C} \\ &= \mathbf{C}^T (\mathbf{C}^T)^{-1} \mathbf{C}^{-1} \mathbf{C} = \mathbf{I} \end{aligned}$$

Damit ergibt sich

$$\rho(\mathbf{u}) = NV(\mathbf{0}, \mathbf{I}) \quad (61)$$

Wird nun $\mathbf{u}$ in seinen Komponenten geschrieben

$$\mathbf{u} = (u_1, u_2, \dots, u_M)^T \quad (62)$$

so gilt für die Komponenten u_i, u_j :

$$\text{cov}\{u_i, u_j\} = 0 \quad (63a)$$

$$\text{var}\{u_i\} = 1 \quad (63b)$$

$$\rho(u_i) = NV(0, 1) \quad (63c)$$

d. h., die Zufallsvariablen sind unkorreliert und damit sind sie auch, weil sie normalverteilt sind, statistisch unabhängig.

$$\rho(u_i, u_j) = \rho(u_i) \rho(u_j) \quad (64)$$

u_i, u_j statistisch unabhängig

Es interessiert die Verteilung von Q_k

$$\begin{aligned}
 Q_k &= (\mathbf{w} - \bar{\mathbf{w}}_k)^T \mathbf{K}_k^{-1} (\mathbf{w} - \bar{\mathbf{w}}_k) \\
 &= \underbrace{(\mathbf{w} - \bar{\mathbf{w}}_k)^T}_{\mathbf{u}^T} \underbrace{\mathbf{C} \mathbf{C}^T}_{\mathbf{u}} (\mathbf{w} - \bar{\mathbf{w}}_k) \\
 &= \mathbf{u}^T \mathbf{u} = |\mathbf{u}|^2 = \sum_{i=1}^M u_i^2
 \end{aligned} \tag{65}$$

Damit ist nach der Verteilung einer Summe von Quadraten von M statistisch unabhängigen normalverteilten Zufallsgrößen mit dem Mittelwert 0 und der Varianz 1 gefragt. Dies ist genau die χ^2 Verteilung $/M/$. Q_k ist damit χ^2 verteilt.

In den Bildern 49 bis 52 sind die gemessenen Verteilungen des Abstandsmaßes Q_k für die Zeichen mit der Kennung "1" und "2" der Handblockschriftprobe H0 als Histogramm dargestellt. Die Dimension des Merkmalsvektors $\mathbf{w}$ ist in Bild 49 und 50 $M = 10$, in den restlichen Diagrammen ist $M = 40$. In die Diagramme mit eingezeichnet ist jeweils die entsprechende χ^2 Verteilung gleicher Dimension.

Man erkennt aus allen Darstellungen, daß die χ^2 Verteilung nur in grober Näherung den gemessenen Verlauf der Dichtefunktion von Q_k wiedergibt. Dabei beschreibt die χ^2 Verteilung in 10 Dimensionen den zugehörigen gemessenen Prozess $\{Q_k\}$ wesentlich besser als die χ^2 Verteilung in 40 Dimensionen. Bei allen Diagrammen fällt das gegenüber der Modellkurve versetzte Maximum der gemessenen Dichtefunktion zu kleineren Werten von Q_k auf. Offensichtlich existieren im M -dimensionalen Merkmalsraum Gebiete, die wesentlich dichter mit Zeichen besetzt sind, als dies von der Normalverteilung erwartet wird. Andererseits sind - und

dies kann besonders bei der Verteilung der Q_k -Maße in 40 Dimensionen festgestellt werden - relativ viele Zeichen mit sehr großen Q_k Werten gezählt worden. Die χ^2 Verteilung erwartet an diesen Stellen wesentlich weniger Zeichen. Hier handelt es sich um Zeichen, deren zugehörigen Punkte im Merkmalsraum sehr weit von dem Klassenschwerpunkt entfernt liegen. Sie würden bei der Modellierung des Prozesses $\{\mathbf{w}/k\}$ durch die Normalverteilung nicht mehr erfaßt werden. Demnach ist nicht sichergestellt, daß ein Klassifikator, der auf der Annahme normalverteilter Merkmalsvektoren $\mathbf{w}$ basiert, diese Zeichen noch richtig klassifizieren kann.

Um diese durch Betrachtung der Bilder 49 bis 52 gewonnenen Erkenntnisse über die Prozesse $\{Q_k\}$ zu bestätigen, wurde - wie in Kapitel 3.3. - der Kolmogoroff-Smirnow-Test durchgeführt. Es gilt dabei festzustellen, ob die Zufallsvariable $\{Q_k(M)\}$ aus einem χ^2 verteilten Prozess entstammt oder nicht. Die Ergebnisse dieses Tests sind in Bild 53 dargestellt.

Dabei ist A wiederum die betragsmäßig größte Abweichung der Modellverteilungsfunktion (χ^2 Verteilung) $F(x)$ von der gemessenen Verteilung $\tilde{F}(x)$. Das obere Diagramm gibt, für jede Klasse k, die Abweichungen A der χ^2 Verteilung mit 10 Freiheitsgraden von der zugehörigen gemessenen Verteilung des Abstandsmaßes Q_k ($M = 10$) wieder. Die Wahrscheinlichkeit dafür, daß bei etwa 2 200 Repräsentanten pro Klasse der Wert $A = 3,4\%$ überschritten wird, ist 1%. Nun liegen sämtliche festgestellten Werte A oberhalb dieser 3,4% Grenze. Die maximale Abweichung beträgt 13,4%, die minimale 3,9%.

Bei der χ^2 Verteilung mit 40 Freiheitsgraden ergeben sich bezüglich der Verteilung des Abstandsmaßes Q_k ($M = 40$) wesentlich größere Abweichungen (unteres Diagramm in Bild 53). Als größte Abweichung A wurde 25,5% gemessen, die kleinste Abweichung beträgt 15,1%. Damit ergibt auch der KS-Test,

daß die Verteilung des Abstandsmaßes Q_k mit $M = 10$ Merkmalen zwar mit Fehler, aber immerhin jedoch wesentlich besser als die Verteilung von Q_k mit 40 Merkmalen durch die χ^2 Verteilung mit M Freiheitsgraden modelliert werden kann.

Die Überprüfung des gegebenen M -dimensionalen Prozesses $\{w/k\}$ hinsichtlich der Übereinstimmung mit der M -dimensionalen Normalverteilung kann ebenfalls anschaulich in der folgenden Weise durchgeführt werden. Wir messen hierzu, für jede Klasse getrennt, den Anteil a der Zeichen außerhalb des Ellipsoides $Q_k = \text{const.}$, wobei der Parameter Q_k von 0 bis $Q_{\max}$ verändert wird. Die Bestimmung von a entspricht einer Integration der gemessenen Wahrscheinlichkeitsdichtefunktion $p(Q_k)$ von Q_k bis ∞ . Dieser Kurve kann nun die entsprechende Kurve des Modellprozesses gegenübergestellt werden. Da bei einem normalverteilten Modellprozess

$$p(w/k) = NV(\bar{w}_k, K_k)$$

die Verteilung des Abstandsmaßes Q_k nach Gleichung (56) eine χ^2 Verteilung ist, ergibt sich der Anteil a der Zeichen, die außerhalb des M -dimensionalen Ellipsoides $Q_k = R^2$ sich befinden, zu

$$a(R) = \int_{Q_k=R^2}^{\infty} p(Q_k) dQ_k, \quad (66)$$

$p(Q_k)$ nach Gleichung (56).

Diese Integrale der χ^2 -Verteilung liegen tabelliert in /A3, B2/ vor. Sie lassen sich nach den folgenden Gleichungen berechnen, wobei eine Normierung der Integrationsgrenze R auf die Varianz des Prozesses $\{u\}$ nach Gleichung (57) durchgeführt wurde.

$$\alpha(RN) = \frac{\Gamma\left(\frac{M}{2}, \frac{Q_k^2}{2}\right)}{\Gamma\left(\frac{M}{2}\right)} \quad (67a)$$

Dabei ist $\Gamma\left(\frac{M}{2}, \frac{Q_k^2}{2}\right)$ die unvollständige Gammafunktion.

$$RN = \frac{R}{\sqrt{\text{var}\{u\}}} = \frac{R}{\sqrt{M}} = \sqrt{\frac{Q_k}{M}} \quad (67b)$$

Speziell für gerade M läßt sich Gleichung (67a) wie folgt schreiben:

$$\alpha(RN) = e^{-\frac{Q_k}{2}} \cdot \sum_{i=0}^{\frac{M}{2}-1} \frac{\left(\frac{Q_k}{2}\right)^i}{i!} \quad (67c)$$

Für ungerades M ergibt sich:

$$\alpha(RN) = \frac{2}{\sqrt{2\pi}} \int_{\frac{Q_k}{2}}^{\infty} e^{-\frac{t^2}{2}} dt + \frac{2}{\sqrt{2\pi}} e^{-\frac{Q_k}{2}} \sum_{r=1}^{\frac{M-1}{2}} \frac{(\sqrt{Q_k})^{2r-1}}{1 \cdot 3 \cdot 5 \cdots (2r-1)} \quad (67d)$$

In dem Diagramm Bild 54 sind die gemessenen Kurven $\alpha(RN)$ der Zeichen mit den Kennungen "1", "2" und "0" abgetragen über der normierten Integrationsgrenze RN. Die Zahl M der zur Berechnung von Q_k verwendeten Merkmale ist dabei $M = 40$. Mit in das Diagramm eingezeichnet sind die durch den Modellprozess Normalverteilung erzeugten Kurven $\alpha(RN)$. Parameter dieser Kurven ist die Dimension M. Aus den Kurven ist zu ersehen, daß der Anteil aller Zeichen außerhalb des Ellipsoides $Q_k = \text{const}$ beim 40-dimensionalen Modell Normalverteilung wesentlich geringer ist als bei den gemessenen Kurven. Die Dichte der Normalverteilung nimmt offenbar wesentlich stärker mit größerem RN ab als die Dichte des Prozesses $\{w/k\}$. Dies wurde bereits bei dem Vergleich der Verteil-

lung von Q_k mit der χ^2 Verteilung festgestellt.

Wir stellen somit eine teilweise recht deutliche Abweichung der gemessenen mit der von der Normalverteilung abgeleiteten Kurve $a(RN)$ fest. Die Ursachen dafür, den Prozess $\{w/k\}$ nicht zufriedenstellend durch eine NV modellieren zu können, kann darin begründet sein, daß entweder das Abklinggesetz der Dichtefunktion $f(Q_k) = e^{-\frac{1}{2}Q_k^2}$ oder das Gesetz, wonach die geometrischen Orte mit konstanter Wahrscheinlichkeitsdichte Ellipsen sind oder daß beide Gesetze nicht zutreffen. Um festzustellen, ob die $a(RN)$ -Kurven einer Modellkurve mit wesentlich schwächerem Abklinggesetz der Dichte mit den gemessenen $a(RN)$ Kurven des Prozesses $\{w/k\}$ besser übereinstimmen, betrachten wir eine weitere Verteilung, die in die Familie des betrachteten Types

$$p(w/k) \sim f(Q_k)$$

nach Gleichung (44) gehört: die Laplaceverteilung.

Der Anteil $a(RN)$ der Zeichen, die außerhalb des M-dimensionalen Ellipsoids $Q_k = R^2$ bei Laplace-verteiltem $\{w/k\}$ sich befindet, läßt sich wiederum durch Integration bestimmen.

Nach Gleichung (57) definieren wir wiederum die Variable u

$$u = C^T(w - \bar{w}_k) .$$

Die M-dimensionale Laplaceverteilung lautet /M2/

$$P(u) = \frac{1}{\sqrt{\pi}^{M-1} \cdot 2 \cdot \alpha^M \sqrt{\frac{M+1}{2}}} e^{-\frac{|u|}{\alpha}} \quad (68a)$$

Nun ist $|u| = \sqrt{Q_k} = Q_k^* = R$. Die Verteilung der Variablen

Q_k^* erhält man durch Integration über die Kugelschale /M2/.

$$P(Q_k^*) = \frac{1}{\alpha^M \Gamma(M)} Q_k^{*M-1} e^{-\frac{Q_k^*}{\alpha}} \quad (69)$$

Der Anteil $a(R)$ ergibt sich zu

$$a(R) = \int_{Q_k^*=R}^{\infty} P(Q_k^*) dQ_k^* \quad (70a)$$

Durch die Substitution $\eta = \frac{Q_k^*}{\alpha}$ erhalten wir ein Integral, dessen Lösung in /B2/ angegeben ist

$$a(R) = \frac{1}{\Gamma(M)} \int_{\eta=R/\alpha}^{\infty} \eta^{M-1} e^{-\eta} d\eta \quad (70b)$$

Mit den Abkürzungen

$$\gamma = M-1 \quad (70c)$$

$$\eta = R/\alpha \quad (70d)$$

erhalten wir

$$a(R) = e^{-\eta} \sum_{i=0}^{\gamma} \frac{\eta^i}{i!} \quad (70e)$$

Wir normieren wieder auf die Varianz des Zufallsprozesses $\{u\}$

$$\text{var} \{u\} = M(M+1) \alpha^2 \quad (70f)$$

Die normierte Integrationsgrenze ist

$$RN = \frac{\eta}{\sqrt{M(M+1)}} = \frac{R}{\sqrt{\text{var}\{u\}}}$$

Somit ergeben sich wieder drei Bestimmungsgleichungen

$$a(RN) = e^{-\eta} \sum_{i=0}^{\eta} \frac{\eta^i}{i!} \quad (70g)$$

$$\eta = M-1 \quad (70h)$$

$$\eta = \sqrt{M(M+1)} \cdot RN \quad (70i)$$

In Bild 55 sind die gemessenen Kurven $a(RN)$ der Zeichen mit der Kennung "0" für verschiedene Dimensionen M (10, 40) über der normierten Integrationsgrenze RN abgetragen. Die entsprechenden Kurven für den Modellprozess sind ebenfalls eingezeichnet.

Aus den Kurven ist ersichtlich, daß die Abweichungen von gemessener und aus dem Modellprozess berechneter Kurve $a(RN)$ bei Laplace-verteiltem Prozess $\{w/k\}$ in $M = 10$ Dimensionen sehr gering sind. Dies ist bei dem 40-dimensionalen Prozess $\{w/k\}$ nicht mehr festzustellen. Auch die Dichte der Laplace-Verteilung nimmt bei größeren RN wesentlich stärker ab als die Dichte des Prozesses $\{w/k\}$.

Abschließend sei noch bemerkt, daß von einer Modellierung des Prozesses $\{w/k\}$ durch eine Laplace-Verteilung - obwohl die Übereinstimmung hinsichtlich der Kurve $a(RN)$ für $M = 10$ relativ gut ist -, wohl abgesehen werden muß, da die eindimensionalen Randverteilungen überhaupt keine Ähnlichkeiten mit Laplace-Verteilungen aufweisen.

4. Realisierung der 2. Stufe

Die 2. Stufe des Erkennungssystems wird durch den eigentlichen Klassifikator realisiert. Er hat die Aufgabe, die durch die Merkmalsextraktionsstufe aufbereitete Information über das zu klassifizierende Zeichen auszuwerten, um dann eine Entscheidung zu treffen, welcher Klasse das Zeichen zuzuordnen ist. Ziel ist dabei eine möglichst fehlerfreie Klas-

sifikation aller Zeichen.

Bei dem Entwurf dieser Stufe wird von dem entscheidungstheoretischen Gesichtspunkt ausgegangen. Dabei wird jede Entscheidung des Klassifikators mit Kosten bewertet. Es geht nun darum, die frei wählbaren Parameter des Klassifikators so einzustellen, daß die entstehenden mittleren Kosten, die als Risiko bezeichnet werden, minimiert werden. Hierbei muß eine genaue Definition der Klassifikationsaufgabe vorliegen. Diese ist gegeben durch eine Stichprobe, in der Zeichen aller zu unterscheidender Klassen enthalten sind. Diese Zeichen sollen den wahren zeichen-erzeugenden Prozess sowohl hinsichtlich Schriftarten wie Schriftqualität möglichst gut repräsentieren.

Zunächst wird die allgemeine Struktur eines derartigen Klassifikators bestimmt. Hierzu bezeichnen wir die Kosten, die bei einer Entscheidung des Klassifikators für Klasse e entstehen, wenn ein Zeichen der Klasse k vorliegt mit

$$C(k, e).$$

Die über ein Zeichen vorliegenden Meßdaten sind hier im Merkmalsvektor $\mathbf{w}$ zusammengefaßt. Dadurch entsteht hier ein Zufallsprozess mit den Variablen $\mathbf{w}$, k und e ,

$$\{\mathbf{w}, k, e\}.$$

Das Risiko ergibt sich somit durch Anwendung des Erwartungswertoperators bezüglich den Variablen $\mathbf{w}$, k , e auf die Kostenfunktion.

$$R_i = \sum_{\mathbf{w}} \sum_k \sum_e C(k, e) p(\mathbf{w}, k, e) \quad (71a)$$

$$R_i = \sum_{\mathbf{w}} \sum_k \sum_e C(k, e) p(e/k, \mathbf{w}) p(k, \mathbf{w}) \quad (71b)$$

Der Erwartungswertoperator wird dabei durch das Zeichen $\sum$

definiert. Dies ist eine symbolische Schreibweise, wie sie auch in /K6/ verwendet wird, und bedeutet, daß bei diskreten Zufallsvariablen zur Bildung des Erwartungswertes summiert, bei kontinuierlichen Variablen integriert werden muß.

Nun hängt die Entscheidung des Klassifikators nur von $\mathbf{w}$ und nicht von k ab, d. h.

$$P(e/k, \mathbf{w}) = P(e/\mathbf{w}) \quad (72)$$

Durch weitere Zerlegung von $p(k, \mathbf{w})$ erhalten wir schließlich

$$R_i = \sum_{\mathbf{w}} \underbrace{\sum_k \sum_e C(k, e) P(e/\mathbf{w}) P(k/\mathbf{w}) \cdot P(\mathbf{w})}_{R_i(\mathbf{w})} \stackrel{!}{=} \text{Min} \quad (73)$$

Ein Minimum für R_i ergibt sich nur dann, wenn für jeden vorliegenden Merkmalsvektor $\mathbf{w}$ $R_i^*(\mathbf{w})$ minimal wird. Dieser minimale Wert für R_i^* kann nun dadurch bestimmt werden, daß sämtliche mögliche Entscheidungen $\tilde{e}$ probeweise getroffen werden und dann der zugehörige Wert für $R_i^*(\mathbf{w})$ bestimmt wird. Dabei sind folgende Entscheidungen möglich

$$\tilde{e} = k \quad 1 \leq k \leq K \quad \text{für eine der } K \text{ Klassen}$$

$$\tilde{e} = 0 \quad \text{Zeichen wird zurückgewiesen}$$

Ist eine Entscheidung $\tilde{e}$ getroffen worden, so ist mit Wahrscheinlichkeit 1 für $\tilde{e}$ entschieden worden, d. h., es ist $p(\tilde{e}/\mathbf{w}) = 1$ (deterministische Entscheidung). Somit ergibt sich als bedingtes Risiko $R_i^*(\tilde{e}, \mathbf{w})$ bei Entscheidung für $\tilde{e}$, wenn ein Vektor $\mathbf{w}$ vorliegt.

$$R_i^*(\tilde{e}, \mathbf{w}) = \sum_k C(k, \tilde{e}) p(k/\mathbf{w}) \quad (74a)$$

Es werden folgende einfache Kostenfunktionen vorgegeben:

richtige Entscheidung: $C(k, e) = 0$; $e = k$ (75a)

Rückweisung : $C(k, e) = C_0$; $e = 0$ (75b)

falsche Entscheidung : $C(k, e) = C_f$; $e \neq k$ (75c)
 $e \neq 0$

Damit ergibt sich für $R_i^*(\tilde{e}, \mathbf{w})$

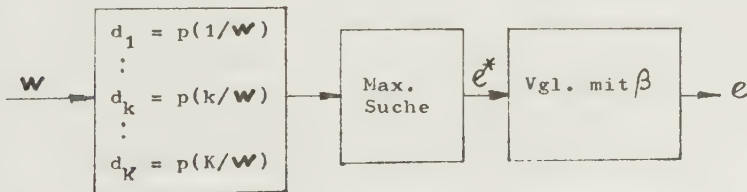
$$R_i^*(\tilde{e}, \mathbf{w}) = \begin{cases} C_0 & \tilde{e} = 0 \\ C_f(1-p(\tilde{e}/\mathbf{w})) & \tilde{e} = k \\ & 1 \leq k \leq K \end{cases} \quad (74b)$$

Führt man schließlich die Abkürzung $\beta = \frac{C_f - C_0}{C_f}$ ein, und ersetzt R_i^* durch $(R_i^*(\tilde{e}, \mathbf{w}) - C_f) \cdot \frac{(-1)}{C_f}$ so muß der Minimumvergleich durch Maximumsuche ersetzt werden, und es ergibt sich schließlich die bekannte Entscheidungsregel /M4/:

Entscheide für Maximum von

$$\begin{cases} \beta \\ p(\tilde{e}, \mathbf{w}) \end{cases} \quad \begin{matrix} \tilde{e} = 0 \\ \tilde{e} = (1, 2, \dots, K) \end{matrix} \quad (76)$$

Die Struktur eines Klassifikators, der ein vorgegebenes Risiko minimiert, ist somit fest gegeben: zu berechnen sind die Rückschlußwahrscheinlichkeiten (a posteriori Wahrscheinlichkeiten).



e^* ist dabei eine vorläufige Entscheidung für eine der Klassen 1, 2, ... K. Sie wird nur dann korrigiert, wenn β größer als der maximale d_k Wert ist.

Die Realisierung eines derartigen Klassifikators erfordert somit die Kenntnis der Rückschlußwahrscheinlichkeiten $p(k/w)$. Diese sind nun im allgemeinen nicht als bekannte Funktionen verfügbar. Eine auch nur teilweise Abspeicherung in Form von Funktionstabellen verbietet der hierzu erforderliche Speicheraufwand. Technisch realisierbare Lösungen erfordern daher eine möglichst gute Schätzung dieser Rückschlußwahrscheinlichkeit durch hantierbare Funktionen.

Im folgenden werden drei Klassifikatoren mit in den Merkmalen w_1 quadratischen Polynomen betrachtet, die, jeder auf eine andere Weise unter Verwendung anderer Annahmen die Rückschlußwahrscheinlichkeiten schätzen.

4.1. Normalverteilungs- und abstands- messender Klassifikator

Nach der Bayesschen Regel kann die Rückschlußwahrscheinlichkeit $p(k/w)$ ausgedrückt werden als Funktion der klassenspezifischen Wahrscheinlichkeitsdichtefunktion $p(w/k)$

$$p(k/w) = \frac{p(w/k) \cdot p(k)}{p(w)} \quad . \quad (77)$$

Der Ausdruck $p(w)$ in Gleichung (77) beeinflusst die Entscheidung des diskutierten Klassifikators, der die entstehenden mittleren Kosten minimiert, und seine Entscheidung nach Gleichung (76) durch Aufsuchen der maximalen Rückschlußwahrscheinlichkeit trifft, nicht. Damit ist zur Berechnung des Entscheidungsmaßes nur die Kenntnis der Wahrscheinlichkeitsdichte $p(w/k)$ und der klassenspezifischen Auftretenswahrscheinlichkeit $p(k)$ erforderlich. Das

Entscheidungsmaß ergibt sich demnach zu

$$d_k = p(\mathbf{w}/k) \cdot p(k) \quad (78)$$

Nun liegen für die klassenspezifischen Wahrscheinlichkeitsdichteverteilungen $p(\mathbf{w}/k)$ aufgrund der Ergebnisse in Kapitel 3 Modellannahmen nahe. Dieses Vorgehen, den Prozess von erzeugten Merkmalen durch Modellfunktionen zu beschreiben, um damit die Entscheidungsmaße zu berechnen, wurde u. a. bereits in den Arbeiten /N1, N2, K1/ gewählt. Aus Vergleichsgründen wird die Ableitung der Entscheidungsgrößen im folgenden nachgezeichnet.

Wir nehmen an, daß der Prozess $\{\mathbf{w}, k\}$ normalverteilt ist. Unter dieser Annahme, ist der Zufallsprozess vollständig durch die klassenspezifische Kovarianzmatrix K_k und den Klassenschwerpunkt $\bar{\mathbf{w}}_k$ beschrieben:

$$p(\mathbf{w}/k) = \frac{1}{\sqrt{(2\pi)^N \det K_k}} \cdot e^{-\frac{1}{2}(\mathbf{w}-\bar{\mathbf{w}}_k)^T K_k^{-1}(\mathbf{w}-\bar{\mathbf{w}}_k)} \quad (79)$$

$$= NV(\bar{\mathbf{w}}_k, K_k)$$

Das Entscheidungsmaß nach Gleichung (78) unter Berücksichtigung von Gleichung (79) ergibt sich nun zu

$$d_k = \frac{p(k)}{\sqrt{(2\pi)^N \det K_k}} \cdot e^{-\frac{1}{2}(\mathbf{w}-\bar{\mathbf{w}}_k)^T K_k^{-1}(\mathbf{w}-\bar{\mathbf{w}}_k)} \quad (80)$$

Wir bezeichnen diesen Klassifikator als Normalverteilungsklassifikator.

Die Entscheidung des Systems wird wiederum nicht beeinflußt, wenn statt d_k nach Gleichung (80) der natürliche Logarithmus von d_k zur Entscheidung herangezogen wird, und gleichzeitig der konstante Faktor $(2\pi)^N$ unberücksichtigt bleibt.

Damit vereinfacht sich die Berechnung der Entscheidungsgröße und ergibt den Ausdruck:

$$d_k = \ln p(k) - \frac{1}{2} \ln \det K_k - \frac{1}{2} [(\mathbf{w} - \bar{\mathbf{w}}_k)^T K_k^{-1} (\mathbf{w} - \bar{\mathbf{w}}_k)] \quad (81)$$

Zu betonen ist, daß die Größe d_k nach Gleichung (81) nun nicht mehr direkt die geschätzte Rückschlußwahrscheinlichkeit $p(k/\mathbf{w})$ selbst ist, sondern nur noch ein Maß für diese Schätzung. Die Leistung des Erkennungssystems, das aufgrund dieses Maßes d_k entscheidet, ist jedoch vollkommen identisch mit demjenigen System, das die Rückschlußwahrscheinlichkeit $p(k/\mathbf{w})$ nach Gleichung (77) unter Berücksichtigung der Modellannahme Gleichung (79) berechnet.

Bei der Realisierung eines Erkennungssystems interessiert, wieviel Koeffizienten zur Berechnung des Entscheidungsmaßes aus den Merkmalen w_i abzuspeichern sind. Für den Normalverteilungsklassifikator sind pro Klasse folgende Größen abzuspeichern:

Größe	Koeffizientenzahl
Klassenkonstante: $\ln p(k) - \frac{1}{2} \ln \det K_k$	1
Schwerpunktvektor $\bar{\mathbf{w}}_k$	M
Klassenspezifische Kovarianzmatrix K_k	$\frac{M}{2} (M + 1)$

Der Koeffizientenbedarf pro Klasse ist damit

$$L = 1 + M + \frac{M}{2}(M + 1) = \left(\frac{M}{2} + \frac{2}{2} \right) \quad (82)$$

Wird nun zusätzlich angenommen, daß die Determinanten der klassenspezifischen Kovarianzmatrizen K_k , sowie die Auftretenswahrscheinlichkeiten $p(k)$ aller Klassen gleich sind, so kann die Entscheidungsregel Gleichung (81) wegen des für alle Klassen gleich großen absoluten Ausdrucks weiter vereinfacht werden. Man erhält damit

$$d_k = - \frac{1}{2} \left[(\mathbf{w} - \bar{\mathbf{w}}_k)^T K_k^{-1} (\mathbf{w} - \bar{\mathbf{w}}_k) \right] \quad (83a)$$

und mit Berücksichtigung von Gleichung (45)

$$d_k = - \frac{1}{2} Q_k \quad (83b)$$

Das Entscheidungsmaß ist damit ausschließlich von dem verallgemeinerten schiefelliptischen Abstandsmaß Q_k nach Gleichung (45) abhängig. Der Klassifikator wird damit als

Abstandsmessender Klassifikator

bezeichnet. Wird bei diesem Klassifikator der Faktor $\frac{1}{2}$ bei der Berechnung von d_k nicht mitberücksichtigt, und gleichzeitig das Vorzeichen des Ausdrucks geändert, so wird die Leistung des Klassifikators nicht beeinflusst, wenn die Maximumsuche durch eine Minimumsuche ersetzt wird, und das Entscheidungsmaß

$$d_k = Q_k \quad (83c)$$

verwendet wird.

Die Anzahl L der pro Klasse abzuspeichernden Koeffizienten ergibt sich hier zu

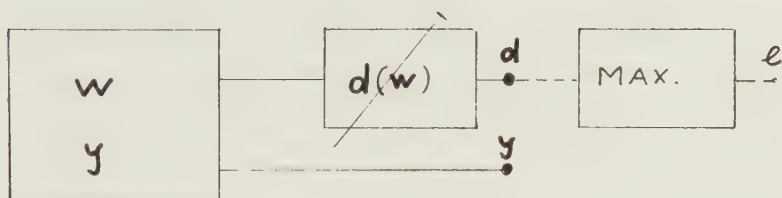
$$L = M + \frac{1}{2} M(M + 1) = \binom{M + 2}{2} - 1 \quad (84)$$

Für beide Klassifikatoren (- Normalverteilungsklassifikator und abstandsmessenden Klassifikator -) ergaben sich aufgrund statistischer Annahmen über den klassenspezifischen Prozess $\{w/k\}$ quadratische Schätzfunktionen. Sind diese Annahmen über den Zufallsprozess nicht exakt zutreffend, und darüber bestehen nach den Ergebnissen von Kapitel 3. keine Zweifel, so ist sicherlich eine Einbuße der Klassifizierungsleistung zu erwarten. In welchem Maße sich dies äußert, wird durch das Experiment festgestellt.

Sicherlich ist die Erkennungsleistung durch besser zutreffende statistische Annahmen noch zu erhöhen. Dies würde jedoch im allgemeinen nicht mehr auf quadratische Schätzfunktionen führen. Gerade die quadratischen Schätzfunktionen sind aber aus Gründen der praktischen Anwendung besonders attraktiv.

4.2. Quadratmittelklassifikator

Im Gegensatz zu den in Kapitel 4.1. betrachteten Klassifikatoren, bei denen sich als Folge von statistischen Annahmen quadratische Schätzfunktionen ergaben, wird in diesem Kapitel ein Klassifikator betrachtet, bei dem der Typus der Schätzfunktion durch ein quadratisches Polynom vorgegeben ist. Eine statistische Hypothese über den Zufallsprozess $\{w/k\}$ wird dabei nicht benutzt. Bei einem derartig konzipierten System kommt es nun darauf an, die Koeffizienten des Schätzpolynoms möglichst geschickt einzustellen:



Wir bestimmen nun die frei verfügbaren Parameter so, daß der mittlere quadratische Fehler zwischen dem Schätzvektor $\mathbf{d}$ und dem nach Gleichung (33a) definierten Zielvektor $\mathbf{y}$ zu einem Minimum wird.

$$S^2 = E \left\{ |\mathbf{y} - \mathbf{d}(\mathbf{w})|^2 \right\} \stackrel{!}{=} \text{Min} \quad (85)$$

(siehe auch Gleichung (34))

Dieser Klassifikator, dessen Koeffizienten nach dem Quadratmittelkriterium optimiert wurden, wird als

Quadratmittelklassifikator

bezeichnet /S1/.

Ergeben sich bei den in Kapitel 4.1. betrachteten Klassifikatoren Abweichungen von der eigentlichen optimalen Schätzfunktion $p(k/\mathbf{w})$, deren Ursache das Nichtzutreffen der statistischen Hypothese ist, so ergeben sich beim Quadratmittelklassifikator mit quadratischen Schätzfunktionen diese Abweichungen wegen der mangelnden Adaptionfähigkeit einer Schätzfunktion von gegebener Form.

Eine unmittelbare Rechtfertigung für die Einführung des Quadratmittelkriteriums zur Optimierung der Koeffizienten besteht darin, daß sich für die Komponenten $d_k(\mathbf{w})$ des Vektors $\mathbf{d}(\mathbf{w})$ genau dann wieder die Rückschlußwahrscheinlichkeiten $p(k/\mathbf{w})$ ergeben, wenn Funktionen von beliebigem Typ und beliebig vielen Parametern zugelassen sind. Dieses Ergebnis ergibt die folgende Ableitung:

Es wird angenommen, daß $\mathbf{d}(\mathbf{w})$ eine Funktion von beliebigem Typ ist und die Forderung von Gleichung (85) erfüllt. Dann muß für alle Abweichungen $\Delta \mathbf{d}(\mathbf{w}) \neq \mathbf{0}$ von der optimalen Lösung gelten

$$s^2(\mathbf{d} + \Delta \mathbf{d}) > s^2(\mathbf{d}) \quad (86)$$

$$\text{für alle } \Delta \mathbf{d} \neq \mathbf{0}$$

Diese Forderung ist global und soll für beliebig große und kleine $\Delta \mathbf{d} \neq \mathbf{0}$ erfüllt werden. Aus Gleichung (85) ergeben sich damit die Größen:

$$\begin{aligned} s^2(\mathbf{d}) &= E\{|\mathbf{y} - \mathbf{d}|^2\} \\ s^2(\mathbf{d} + \Delta \mathbf{d}) &= E\{|\mathbf{y} - \mathbf{d} - \Delta \mathbf{d}|^2\} \\ &= E\{|\mathbf{y} - \mathbf{d}|^2\} - 2E\{\Delta \mathbf{d}^T [\mathbf{y} - \mathbf{d}]\} \\ &\quad + E\{|\Delta \mathbf{d}|^2\} \end{aligned}$$

In die Ungleichung Gleichung (86) eingesetzt, ergibt sich:

$$0 < E\{|\Delta \mathbf{d}|^2\} + 2E\{\Delta \mathbf{d}^T [\mathbf{d} - \mathbf{y}]\}, \Delta \mathbf{d} \neq \mathbf{0} \quad (87)$$

Da der Term $E\{|\Delta \mathbf{d}|^2\}$ stets größer Null ist, ist Gleichung (87) sicher für alle $\Delta \mathbf{d} \neq \mathbf{0}$ erfüllt, wenn der zwei-

te Ausdruck verschwindet:

$$E\{\Delta \mathbf{d}^T [\mathbf{d} - \mathbf{y}]\} = 0 \quad (88a)$$

Der Erwartungswertoperator kann nun aufgelöst werden, indem über alle vorkommenden $\mathbf{w}$ und $\mathbf{y}$ unter Berücksichtigung der entsprechenden Wahrscheinlichkeiten summiert wird:

$$\sum_{\mathbf{w}} \sum_{\mathbf{y}} \Delta \mathbf{d}^T [\mathbf{d} - \mathbf{y}] p(\mathbf{w}, \mathbf{y}) = 0 \quad (88b)$$

$$\underbrace{\hspace{10em}}_{= p(\mathbf{y}/\mathbf{w}) \cdot p(\mathbf{w})}$$

$$\rightarrow \sum_{\mathbf{w}} \sum_{\mathbf{y}} \Delta \mathbf{d}^T [\mathbf{d} - \mathbf{y}] p(\mathbf{y}/\mathbf{w}) \cdot p(\mathbf{w}) = 0 \quad (88c)$$

$$\rightarrow \sum_{\mathbf{w}} \Delta \mathbf{d}^T \underbrace{\sum_{\mathbf{y}} [\mathbf{d} - \mathbf{y}] p(\mathbf{y}/\mathbf{w}) \cdot p(\mathbf{w})}_{\mathcal{F}} = 0 \quad (88d)$$

Gleichung (88d) ist dann, und nur dann für beliebige $\Delta \mathbf{d} \neq 0$ erfüllt, wenn der Ausdruck $\mathcal{F}$ verschwindet:

$$\sum_{\mathbf{y}} [\mathbf{d} - \mathbf{y}] p(\mathbf{y}/\mathbf{w}) = 0 \quad (89a)$$

Durch Ausmultiplizieren erhalten wir:

$$\sum_{\mathbf{y}} \mathbf{d} p(\mathbf{y}/\mathbf{w}) = \sum_{\mathbf{y}} \mathbf{y} p(\mathbf{y}/\mathbf{w}) \quad (89b)$$

Nun hängt $\mathbf{d}$ nur von $\mathbf{w}$ ab und kann damit vor das Summenzeichen geschrieben werden:

$$\mathbf{d} \sum_{\mathbf{y}} p(\mathbf{y}/\mathbf{w}) = \sum_{\mathbf{y}} \mathbf{y} p(\mathbf{y}/\mathbf{w}) \quad (89c)$$

Nach Definition ist aber $\sum_{\mathbf{y}} p(\mathbf{y}/\mathbf{w}) = 1$, und damit ergibt sich:

$$d = \sum_y y p(y/w) = E\{y/w\} \quad (90)$$

Bei uneingeschränkter Optimierung der Schätzfunktionen erhalten wir nach Gleichung (90) die Regressionsfunktion /R2, Seite 264/. Diese Funktion $d(w)$ erfüllt Gleichung (86), und zwar für beliebig große und kleine $\Delta d \neq 0$. Jede andere Abbildung $d^*(w)$ führt deshalb notwendig zu einem größeren $S^2(d^*)$. Damit ist die Lösung der Minimaufgabe eindeutig. Die Komponenten von d sind, wegen der besonderen Definition von y nach Gleichung (33a), die Rückschlußwahrscheinlichkeiten:

$$d_k(w) = E\{y_k/w\} = p(k/w) \quad (91)$$

Wir erhalten damit für diesen Grenzfall wieder die geforderte Struktur des Erkennungssystems, dessen Entscheidungsregel in Gleichung (76) definiert ist.

Schränkt man die Optimierung durch Vorgabe des Types von d_k jedoch ein - wie dies im folgenden getan wird - indem nur quadratische Funktionen betrachtet werden, so können die Komponenten von d interpretiert werden als geschätzte Werte für die Rückschlußwahrscheinlichkeiten $p(k/w)$. Zur Optimierung nach Gleichung (85) werden vollständige quadratische Ansatzpolynome

$$d(w) = A^T x(w) \quad (92)$$

mit

$$A = (a_1, a_2, \dots, a_K) \quad (93a)$$

$$x = (1, w_1, w_2, \dots, w_i, \dots, w_M, \dots, w_i w_m, \dots, w_i w_M)^T \quad (93b)$$

verwendet. A wird dabei als Koeffizientenmatrix, x als

Eigenschaftsvektor bezeichnet, dessen Komponenten die Merkmale w_i sowie Produkte der Merkmale $w_i w_j$ sind. Nach Gleichung (92) und Gleichung (93) läßt sich die Komponente d_k wie folgt berechnen:

$$d_k = \mathbf{a}_k^T \mathbf{x}(\mathbf{w}) \quad (94)$$

Zur Berechnung der Schätzung d_k der Rückschlußwahrscheinlichkeit müssen sämtliche Koeffizienten von $\mathbf{a}_k$ verfügbar sein. Dies sind beim vollständig quadratischen Schätzpolynom (Ansatzpolynom)

$$L = L_{VQ} = \binom{M + 2}{2} \quad (95)$$

Koeffizienten. Damit ist die Zahl der pro Klasse benötigten Koeffizienten beim vollständig quadratischen Schätzpolynom des Quadratmittelklassifikators genauso groß wie beim Normalverteilungsklassifikator.

Die Optimierung der Koeffizientenmatrix $\mathbf{A}$ nach Gleichung (85) unter Berücksichtigung von Gleichung (92) erfolgt durch Lösung des in den Variablen x_i linearen Gleichungssystems

$$E\{\mathbf{x} \mathbf{x}^T\} \mathbf{A} = E\{\mathbf{x} \mathbf{y}^T\} \quad (96)$$

/S1/. In Kapitel 2.1. (Gleichung (33) bis (37)) wurde bereits abgeleitet, welche Koeffizienten sich für den Fall des in den Merkmalen w_i linearen Schätzpolynoms ergeben. In dem vorliegenden Fall des quadratischen Polynoms sind keine weiteren Vereinfachungen des Gleichungssystems mehr möglich. Wird dieses System linearer Gleichungen durch die Anwendung des Gauß-Jordan-Verfahrens aufgelöst, bei dem schrittweise eine Einheitsmatrix erzeugt wird, so kann der Ablauf des Prozesses hinsichtlich des Optimierungszieles S^2 günstig gesteuert werden /S1/. Dies ist möglich, da bei der Berechnung der $\mathbf{A}$ -Matrix eine Reihenfolge nicht vorge-

geschrieben ist, in der die Zeilen der Matrix auf Normalform gebracht werden. Im Hinblick darauf, das Optimierungskriterium, nämlich den Endwert S^2 , möglichst frühzeitig zu erreichen, wählen wir wieder die MS-Strategie, die bereits beim linearen Klassifikator angewandt wurde, und dort zur Ermittlung wichtiger Merkmale diente. Diese Strategie wird für den vorliegenden Fall quadratischer Schätzpolynome mit korrelierten Variablen x_i kurz erläutert.

Hierzu betrachten wir einen Zwischenzustand in der Bearbeitung der Matrix, in den bereits $L < L_{VQ}$ Spaltenvektoren auf Normalform gebracht worden sind. Dies bedeutet, daß die zu diesen Spaltenvektoren gehörenden Variablen x_i bereits zur Schätzung d_k der Rückschlußwahrscheinlichkeit $p(k/\mathbf{w})$ verwendet wurden. Die zugehörigen optimierten Koeffizienten sind ebenfalls verfügbar. Damit wird das Optimierungskriterium nach Gleichung (85) erfüllt, wenn der Vektor $\mathbf{x}$ in Gleichung (92) ersetzt wird durch einen Vektor $\mathbf{x}^*$, dessen Komponenten die bereits berücksichtigten L Variablen x_i sind. Die erreichte Reststreuung S_L^2 ist im allgemeinen größer als der erreichte Endwert, kann aber wie in /S1/ gezeigt wird, nicht kleiner als dieser sein.

Wird nun ein neuer Spaltenvektor auf Normalform gebracht, so wird den bereits berücksichtigten L Variablen eine neue hinzugefügt. Das aktuell gültige Schätzpolynom besitzt nun die Länge $L + 1$. Die zugehörigen Koeffizienten des Polynoms sind, und dies ist ein wesentlicher Unterschied zu dem Koeffizientensatz des linearen Klassifikators mit den unkorrelierten Variablen w_i , vollständig geändert worden. Die aktuelle Reststreuung hat sich in diesem Zustand auf den Wert S_{L+1}^2 verkleinert.

Da wir in der Wahl der neu zu berücksichtigenden Variablen frei sind, wählen wir nun diejenige, die die erreichte Reststreuung um den größten Betrag ΔS^2 zu verringern vermag

(MS-Strategie). Hierzu müssen sämtliche Beiträge ΔS_i^2 aller noch verfügbaren Variablen in jedem Bearbeitungszustand neu berechnet werden, da diese davon abhängig sind, welche Variablen bereits zur Schätzung d_k vorgesehen sind. Dies bedeutet, daß lineare Abhängigkeiten vollständig berücksichtigt werden können. Variable, die bereits durch die ausgesuchten Variablen linear darstellbar sind, werden an dem Verschwinden ihres Hauptdiagonalelements erkannt und eliminiert. Mit dieser Strategie wird versucht, die kleinste Reststreuung S_L^2 zu erreichen, die mit L Variablen bei vorgegebener Struktur und Länge L_{ANS} des Ansatzpolynoms erreicht werden kann. Wie in /S1/ gezeigt wird, können zwar Fälle konstruiert werden, in der diese Methode nicht bei jedem $L < L_{VQ}$ die kleinstmögliche Reststreuung S_L^2 liefert, Experimentelle Untersuchungen mit realistischen Stichproben /S6/ zeigen jedoch, daß diese Strategie im numerischen Sinne optimal ist.

Da wir den Regressionsprozess stets vorzeitig abbrechen können, ist die Anzahl der zur Berechnung von d_k notwendigen Koeffizienten bei vorgegebener Zahl M von verwendeten Merkmalen w_i zusätzlich einstellbar, wobei die maximale Größe von L durch Gleichung (95) definiert ist:

$$L < L_{VQ} = \binom{M + 2}{2} \quad \begin{array}{l} \text{beim nicht vollständig} \\ \text{quadr. Schätzpolynom} \end{array}$$

Die zur Schätzung benutzten Variablen x_i stehen am vorderen Ende der MS-Rangordnungsliste. Die vollständige MS-Rangordnungsliste ermöglicht dabei rückblickend interessante Zusammenhänge aufzuzeigen; es kann festgestellt werden, ob die Rangreihenfolge unter den Merkmalen w_i auch Schlüsse auf die Rangreihenfolge unter den Einflußgrößen x_i , d. h., den Merkmalsverknüpfungen zuläßt.

4.3. Vergleichende Gegenüberstellung der quadratischen Klassifikatoren

In Kapitel 4.1. und 4.2. wurde je ein Verfahren zur Berechnung eines Klassifikators angegeben. Beim Verfahren nach Kapitel 4.1. wurde eine statistische Hypothese über den Prozess $\{\mathbf{w}/k\}$ eingeführt: dieser Zufallsprozess soll durch eine Normalverteilung gut beschreibbar sein. Daraus ergibt sich unmittelbar sowohl die Struktur wie die Größe der Koeffizienten der Schätzfunktion des Klassifikators, welcher das Risiko minimiert. Die Schätzfunktion zur Berechnung des Entscheidungsmaßes d_k ist ein in den Merkmalen w_1 vollständig quadratisches Polynom.

Beim Verfahren nach Kapitel 4.2. wurde die Struktur der Schätzfunktion durch ein vollständig quadratisches Polynom vorgegeben. Zur Berechnung der Koeffizienten wurde das Quadratmittelkriterium herangezogen.

Es interessiert nun, in welcher Beziehung diese beiden Klassifikatoren stehen, deren Entscheidungsmaß mit quadratischen Polynomen berechnet werden, die jedoch nach verschiedenen Methoden bestimmt wurden.

Zunächst muß nochmals betont werden, daß die Entscheidungsmaße d_k nach Gleichung (81) (bzw. Gleichung (83)) beim Normalverteilungsklassifikator (bzw. abstandsmessenden Klassifikator) nicht die geschätzten Rückschlußwahrscheinlichkeiten selbst sind, sondern nur Maße dafür. Deren Verwendung garantiert jedoch die gleiche Erkennungsleistung des Systems, wie wenn die Rückschlußwahrscheinlichkeiten $p(k/\mathbf{w})$ selbst zur Entscheidung herangezogen werden würden. Die Rückschlußwahrscheinlichkeiten selbst sind nach Gleichung (77) und (79) definiert. Daraus ist ersichtlich, daß $p(k/\mathbf{w})$ selbst kein quadratisches Polynom in w_1 ist, sondern eine e-Funktion mit dem quadratischen Polynom im Exponenten.

Bei dem Verfahren nach Kapitel 4.2. werden jedoch die Rückschlußwahrscheinlichkeiten direkt durch das quadratische Polynom nach Gleichung (92) (bzw. Gleichung (94)) geschätzt. Zwar können sich wegen der Quadratmittelloptimierung Schätzwerte für die Rückschlußwahrscheinlichkeiten $p(k/\mathbf{w})$ ergeben, die größer 1 und kleiner 0 sind, Ziel der Schätzung ist jedoch, mit Wahrscheinlichkeit 1 die wahre Klasse, und mit Wahrscheinlichkeit 0 die fremden Klassen eines Zeichens zu schätzen. Dieses Ziel wird im Grenzfall der uneingeschränkten Optimierung von Schätzfunktionen beliebiger Struktur exakt erreicht: die Rückschlußwahrscheinlichkeiten werden fehlerfrei bestimmt.

Unter Beachtung dieser Sachverhalte wird deutlich, daß der Quadratmittelklassifikator, dessen Schätzfunktion ein quadratisches Polynom ist, nicht zum Normalverteilungsklassifikator werden kann. Das heißt, es ergeben sich stets verschiedene Koeffizienten für die Schätzpolynome nach Gleichung (81) (bzw. (83)) und Gleichung (94).

Liegt ein exakt normalverteilter Prozess $\{\mathbf{w}/k\}$ vor, so ist der Normalverteilungsklassifikator in der Lage, mit den quadratischen Schätzfunktionen eine optimale, durch kein anderes System zu überbietende Klassifikation durchzuführen. Hier reichen quadratische Funktionen, die nur Maße für die Rückschlußwahrscheinlichkeiten $p(k/\mathbf{w})$ sind, zur optimalen Entscheidung aus. Der Quadratmittelklassifikator liefert unter diesen speziellen Bedingungen wiederum auch andere Schätzpolynome und vermag daher dieses optimale Klassifizierungsergebnis nicht zu erreichen. Der Grund dafür ist, daß mit quadratischen Polynomen die Rückschlußwahrscheinlichkeiten $p(k/\mathbf{w})$ bei normalverteiltem $\{\mathbf{w}/k\}$ nicht exakt zu bestimmen sind. Erst die Zulassung beliebiger Schätzfunktionen, die nach dem quadratischen Mittel optimiert werden, würde exakt die Berechnung der Rückschlußwahrscheinlichkeit $p(k/\mathbf{w})$ ergeben. Dies ist in diesem

Fall kein Polynom, sondern die durch Gleichung (80) definierte Funktion.

Trotz dieser aufgezeigten Unterschiede der Klassifikatoren existieren untereinander gewisse Beziehungen und Ähnlichkeiten. Um diese aufzuzeigen, werden die zur Entscheidung verwendeten Größen d_k auf eine einheitliche Form gebracht. Beim Quadratmittelklassifikator werden, um dies zu erreichen, nur vollständig quadratische Schätzpolynome betrachtet.

Die in den Merkmalen w_i vollständig quadratischen Polynome lassen sich stets in der Form

$$d_k = \mu_k - (\mathbf{w} - \tilde{\mathbf{w}}_k)^T \tilde{\mathbf{G}}_k (\mathbf{w} - \tilde{\mathbf{w}}_k) \quad (97)$$

darstellen. Der Ausdruck μ_k ist eine klassenspezifische Konstante, der folgende Ausdruck in der Gleichung ist eine quadratische Form mit $\tilde{\mathbf{G}}_k$ als Kernmatrix. Wir interessieren uns nun für die Gestalt der Gebiete im Merkmalsraum, auf denen die Maße für die geschätzten Rückschlußwahrscheinlichkeiten d_k konstant sind. Hierzu führen wir in Anlehnung an Gleichung (49) - (54) folgende Größen ein:

$$\tilde{d}_k = \mu_k - d_k \quad (98)$$

$$\tilde{\mathbf{z}} = \tilde{\mathbf{B}}^T (\mathbf{w} - \tilde{\mathbf{w}}_k) \quad (99)$$

$$\tilde{\mathbf{L}}_k = \begin{pmatrix} \lambda_1 & \dots & 0 \\ 0 & & \lambda_M \end{pmatrix} \quad (100)$$

Hierbei ist $\tilde{\mathbf{B}}$ die Matrix der Eigenvektoren von $\tilde{\mathbf{G}}_k$, $\tilde{\mathbf{L}}_k$ eine Diagonalmatrix, deren Hauptdiagonalelemente die Eigenwerte von $\tilde{\mathbf{G}}_k$ sind. Damit muß dann gelten:

$$\tilde{\mathbf{B}}^T \tilde{\mathbf{G}}_k \tilde{\mathbf{B}} = \tilde{\mathbf{L}}_k \quad (101a)$$

$$\tilde{\mathbf{B}}^T \tilde{\mathbf{B}} = \mathbf{I} \quad (101b)$$

$$\tilde{\mathbf{G}}_k = \tilde{\mathbf{B}} \tilde{\mathbf{L}}_k \tilde{\mathbf{B}}^T \quad (101c)$$

Gleichung (97) ist nun auf die Form gebracht:

$$\tilde{d}_k = (\mathbf{w} - \tilde{\mathbf{w}}_k)^T \tilde{\mathbf{B}} \tilde{\mathbf{L}}_k \tilde{\mathbf{B}}^T (\mathbf{w} - \tilde{\mathbf{w}}_k) \quad (102a)$$

$$= \tilde{\mathbf{z}}^T \tilde{\mathbf{L}}_k \tilde{\mathbf{z}} \quad (102b)$$

$$\tilde{d}_k = \tilde{z}_1^2 \lambda_1 + \tilde{z}_2^2 \lambda_2 + \dots \dots \tilde{z}_M^2 \lambda_M \quad (102c)$$

Entscheidend für die Gestalt des Gebietes $\tilde{d}_k = \text{const}$ sind die Vorzeichen der Eigenwerte λ_i . Besitzen diese einheitliches Vorzeichen ($\tilde{\mathbf{G}}_k$ definit), so entstehen Hyperellipsoide. Treten verschiedene Vorzeichen für λ_i auf, so existieren Ebenen mit Ellipsen und Ebenen mit Hyperbeln als Schnittfiguren. Der Mittelpunkt der Kegelschnitte ist dabei stets $\tilde{\mathbf{w}}_k$.

Nun ist aus Gleichung (81) und Gleichung (83) unmittelbar ersichtlich, daß diese Gleichungen, die die Entscheidungsmaße des Normalverteilungsklassifikators und des abstandsmessenden Klassifikators definieren, bereits in der gewünschten allgemeinen Form vorliegen. Eine Umformung ist dagegen für die Gleichung, die das Entscheidungsmaß d_k des Quadratmittelklassifikators definiert (Gleichung (94)), notwendig. Nach Voraussetzungen werden nur vollständig quadratische Polynome betrachtet. Damit ist in Gleichung (94)

$$\mathbf{x} = (\mathbf{w}_0, \mathbf{w}_1, \dots, \mathbf{w}_M, \dots, \mathbf{w}_i \mathbf{w}_j, \dots, \mathbf{w}_M \mathbf{w}_M)^T \quad (103a)$$

$$\downarrow$$

$$= 1 \quad i = 1, \dots, M \quad (103b)$$

$$j = i, \dots, M \quad (103c)$$

Nun kann aus den Komponenten von $\mathbf{a}_k$ (Gleichung (94)) eine Matrix $\mathbf{A}_k^*$ derart aufgebaut werden, daß sich $\mathbf{d}_k$ schreiben läßt in der Form

$$\mathbf{d}_k = (\mathbf{1} \ \mathbf{w})^T \mathbf{A}_k^* \begin{pmatrix} 1 \\ \mathbf{w} \end{pmatrix} \quad (104)$$

Hierzu werden sämtliche Komponenten von $\mathbf{a}_k$, bis auf die zu den Variablen w_i und w_o gehörigen halbiert. Mit diesen Elementen wird nun eine untere Dreiecksmatrix besetzt. Durch Spiegelung ergibt sich dann die gesuchte Matrix $\mathbf{A}_k^*$. Um schließlich die gewünschte Form von $\mathbf{d}_k$ zu erhalten, teilen wir $\mathbf{A}_k^*$ wie folgt auf:

$$\mathbf{A}_k^* = \begin{pmatrix} h_k & h_k^T \\ h_k & H_k \end{pmatrix} \quad (105)$$

Durch Einsetzen von Gleichung (104) in Gleichung (105) ergibt sich die Darstellung:

$$\mathbf{d}_k = \mathbf{g}_k - (\mathbf{w} - \mathbf{w}_{ok})^T \mathbf{G}_k (\mathbf{w} - \mathbf{w}_{ok}) \quad (106a)$$

$$\text{mit} \quad \mathbf{g}_k = h_k + \mathbf{w}_{ok}^T h_k \quad (106b)$$

$$\mathbf{w}_{ok} = -\mathbf{H}_k^{-1} h_k = \mathbf{G}_k^{-1} h_k \quad (106c)$$

$$\mathbf{G}_k = -\mathbf{H}_k \quad (106d)$$

Die Entscheidungsmaße aller drei Klassifikatoren sind damit auf die mit Gleichung (97) definierte Form gebracht worden. Die Parameter der Gleichung sind für die jeweiligen Systeme verschieden und wie folgt zu berechnen:

Normalverteilungsklassifikator:

$$\mu_k = \ln p(k) - \frac{1}{2} \ln \det \mathbf{K}_k \quad (107a)$$

$$\tilde{\mathbf{w}}_k = \overline{\mathbf{w}}_k \quad (107b)$$

$$\tilde{\mathbf{G}}_k = \frac{1}{2} \mathbf{K}_k^{-1} \quad (107c)$$

Abstandsmessender Klassifikator:

$$u_k = 0 \quad (108a)$$

$$\tilde{\mathbf{w}}_k = \overline{\mathbf{w}}_k \quad (108b)$$

$$\tilde{\mathbf{G}}_k = \frac{1}{2} \mathbf{K}_k^{-1} \quad (108c)$$

Quadratmittelklassifikator:

$$u_k = g_k \quad (109a)$$

$$\tilde{\mathbf{w}}_k = \mathbf{w}_{ok} \quad (109b)$$

$$\tilde{\mathbf{G}}_k = \mathbf{G}_k \quad (109c)$$

Die Kernmatrix für Normalverteilungsklassifikator und abstandsmessenden Klassifikator ist die Inverse der Kovarianzmatrix $\mathbf{K}_k$, und damit positiv semidefinit. Sämtliche Eigenwerte von $\mathbf{K}_k^{-1}$ sind größer oder gleich 0. Als Gebiete mit konstanten d_k Werten ergeben sich damit Hyperellipsoide. Die Lage und Richtung dieser geometrischen Orte wurde für das Beispiel eines zweidimensionalen Merkmalsraumes, in dem die Zeichen nur durch zwei Merkmale w_i beschrieben werden, für Normalverteilungsklassifikator und abstandsmessenden Klassifikator bereits in Kapitel 2.3. dargestellt. Es wurden dort die geometrischen Orte der quadratischen Form $Q_k = \overline{Q}_k = M$ gezeichnet, die ebenfalls Orte mit konstanten d_k sind (Bild 45 bis 48). Der zugehörige d_k -Wert ist beim Normalverteilungsklassifikator von weiteren klassenspezifischen Parametern abhängig:

$$d_k = \ln p(k) - \frac{1}{2} \ln \det K_k - \frac{1}{2} \bar{q}_k \quad (110)$$

während beim abstandsmessenden Klassifikator die Zuordnung gilt:

$$d_k = - \frac{1}{2} \bar{q}_k \quad (111)$$

Über die Eigenschaften der Kernmatrix G_k des Quadratmittelklassifikators kann von vornherein keine Aussage gemacht werden. G_k kann hier positiv bzw. negativ definit oder auch nicht definit sein. Um dies festzustellen, muß ein Experiment durchgeführt werden, in dem G_k explizit berechnet wird. Da sich die Matrixelemente von G_k unmittelbar aus den Komponenten von a_k , die zu den Variablen $w_i w_j$ gehören, ergeben, muß gelten:

$$\sum_{k=1}^K G_k = 0 \quad (112)$$

Um dies zu zeigen, gehen wir von folgender Eigenschaft der Matrix A , die Gleichung (96) erfüllt, aus:

$$Ae = (a_1, a_2, \dots, a_k)e = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad (113a)$$

$$\text{dabei ist } e = (1, 1, \dots, 1)^T. \quad (113b)$$

Bei der $L \times K$ Matrix A ist somit $\text{Dim } [e] = K$ und $\text{Dim } [(1 \ 0)^T] = L$

Die Gültigkeit von Gleichung (113a) kann nachgewiesen werden, indem die Berechnungsvorschrift für A nach Gleichung (96) eingesetzt wird. Man erhält dann:

$$Ae = [E\{xx^T\}]^{-1} \cdot E\{xy^T\} e = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad (114a)$$

Wenn Gleichung (114a) erfüllt ist, dann muß auch gelten:

$$E\{xy^T\} \cdot e \stackrel{!}{=} E\{xx^T\} \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad (114b)$$

Spalten wir nun den Vektor x entsprechend Gleichung (93b) auf

$$x^T = (1, x^{*T}) \quad (114c)$$

und setzen diese Gleichung in Gleichung (114b) ein, dann erhalten wir:

$$\overline{xy^T} \cdot e \stackrel{!}{=} \begin{pmatrix} 1 & \overline{x^{*T}} \\ \frac{1}{x^*} & \overline{x^* x^{*T}} \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ \frac{1}{x^*} \end{pmatrix} \quad (114d)$$

Wegen der Definition von y (Gleichung 33a) gilt:

$$\overline{xy^T} = (p_1 \overline{x}_1, p_2 \overline{x}_2, \dots, p_K \overline{x}_K) \quad (114e)$$

wobei p_k die Auftretenswahrscheinlichkeit der Klasse k ist. Mit Gleichung (114e) kann $\overline{xy^T} \cdot e$ geschrieben werden:

$$\overline{xy^T} \cdot e = \sum_{k=1}^K p_k \overline{x}_k = \overline{x} = \begin{pmatrix} 1 \\ \frac{1}{x^*} \end{pmatrix} \quad (114f)$$

Damit ist die Gültigkeit von (114d) und somit auch von Gleichung (113a) nachgewiesen.

Gleichung (113a) besagt, daß sämtliche Zeilensummen der A Matrix, bis auf die erste, verschwinden. Nach Gleichung (106) sind die Hauptdiagonalelemente von G_k unmittelbar Elemente von a_k , die übrigen Elemente von G_k ergeben sich durch Halbierung der Komponenten von a_k . Die Summierung aller K Matrizen G_k ist damit identisch mit der Summierung der Matrixzeilen von A und ergibt demnach die Nullmatrix, da Elemente der ersten Matrixzeile von A nicht in G_k erscheinen.

Der Mittelpunkt der quadratischen Form ist durch den Vektor $\mathbf{w}_{0k}$ definiert. Für $\mathbf{w} = \mathbf{w}_{0k}$ ergibt sich die Schätzung d_k der Klassenzugehörigkeit zu $d_k = g_k$. Damit kann bei Kenntnis von g_k unmittelbar auf Lage des Mittelpunkts im Merkmalsraum geschlossen werden. Ist $g_k \approx 1$, d. h., die geschätzte Rückschlußwahrscheinlichkeit für die Klasse k ist ungefähr 100%, so muß der Mittelpunkt $\mathbf{w}_{0k}$ im Gebiet der Klasse k liegen.

Aus Gleichung (97) kann über den Wertebereich von d_k nur dann eine Aussage gemacht werden, wenn festgestellt wurde, daß $\mathbf{G}_k$ definit ist. Dann gilt:

$$d_k \leq g_k, \quad \text{wenn } \mathbf{G}_k \text{ positiv definit}$$

$$d_k \geq g_k, \quad \text{wenn } \mathbf{G}_k \text{ negativ definit}$$

Für diesen Fall sind die $g_k = \text{const}$ Gebiete Hyperellipsoide. Dem Mittelpunkt eines jeden Hyperellipsoides ist der Wert $d_k = g_k$ zugeordnet.

Um eine anschauliche Vorstellung darüber vermitteln zu können, welche Kurven konstanter Schätzung d_k bei der Optimierung der Koeffizienten im quadratischen Mittel sich ergeben, beschränken wir uns auf einen darstellbaren zweidimensionalen Raum. Wir betrachten wieder Ebenen im Merkmalsraum, die von zwei Basisvektoren $\mathbf{b}_i, \mathbf{b}_j$ aufgespannt werden. Damit sind die Ergebnisse unmittelbar vergleichbar mit den Darstellungen in den Bildern 45 bis 48, in denen die Klassengebiete durch die Ellipsen $Q_k = 2$ beschrieben werden, und die für abstandsmessenden und Normalverteilungsklassifikator auch Kurven mit konstantem Entscheidungsmaß d_k sind.

Zur Berechnung der Koeffizientenmatrix $\mathbf{A}$ des Quadratmittelklassifikators verwenden wir wieder die Merkmalsvektoren $\{\mathbf{w}\}$ der gesamten Stichprobe H_0 . Jedes Zeichen wird dabei durch den zweidimensionalen Vektor

$$\mathbf{w} = (w_i, w_j)^T$$

beschrieben, wobei die Indizes i, j die Rangnummern des Merkmals entsprechend der Größe des zugehörigen Eigenwertes sind. Entsprechend den Darstellungen in Bild 45 bis 47 wird gewählt:

i	j	Ebene
1	2	w_1/w_2
1	3	w_1/w_3
1	10	w_1/w_{10}

Damit ergeben sich drei Koeffizientenmatrizen, wobei jede für eine Ebene zuständig ist. Aus diesen Matrizen werden nun die zweidimensionalen Kernmatrizen G_k berechnet, und die entstehenden $d_k = \text{const}$ Kurven in den Bildern 56 bis 58 gezeichnet.

Um die Darstellungen nicht zu überlasten, wurden nur die Kurven konstanter Schätzung d_k für die Kennungen "0", "1", "2" eingetragen. Um die Klassengebiete zu kennzeichnen, wurden die zu den Zeichen gehörigen Merkmale w_i, w_j mit in das Koordinatensystem eingezeichnet, und zwar für einen statistischen Extrakt von 500 Zeichen pro Klasse. Die Ortskurven von konstanten d_k der Zeichen der Kennung "1" und "2" in der $w_1 w_2$ Ebene (Bild 56) sind Hyperbeln, die entsprechenden Kurven für die Kennung "0" Ellipsen. Die Halbachsen der Ellipsen werden für abnehmendes d_0 kleiner. Dem Mittelpunkt der Ellipse ist der Wert $d_0 = g_0 = -0,038$ zugeordnet, der gleichzeitig eine untere Grenze für d_0 darstellt. Weiter ist erkennbar, daß die 100% Schätzung ($d_2 = 1$) der Klasse "2" durch das Klassengebiet der 2 verläuft, wobei der zugehörige symmetrische Ast dieser Hyperbel in einem Bereich

liegt, in dem keine Meßwerte von Zeichen liegen. Gleichzeitig verlaufen die Hyperbeläste $d_2 = 0$ durch die Punktmenge der Zeichen der Klasse "0" und "1". Das Ziel der Schätzung, die eigene Klasse mit der Wahrscheinlichkeit 100%, die übrigen mit 0% zu schätzen, wird, zumindest für die Trennung dieser drei betrachteten Klassen auch bei diesen kleinen Schätzgleichungslängen relativ gut erreicht.

Aus den Bildern 57 und 58 geht ebenfalls hervor, daß sich auch in diesen Ebenen Hyperbeln und Ellipsen als Orte konstanter Schätzung d_k ergeben.

Im Gegensatz zu den geometrischen Orten von konstantem Entscheidungsmaß des Normalverteilungsklassifikators und abstandsmessenden Klassifikators, können die $d_k = \text{const}$ Kurven des Quadratmittelklassifikators nicht als Orte aufgefaßt werden, die ein Klassengebiet umschreiben. Infolgedessen sind die Darstellungen in Bild 45 bis 47 optisch eindrucksvoller. Ein weiterer wesentlicher Unterschied zwischen Quadratmittelklassifikator und den beiden anderen betrachteten Klassifikatoren besteht darin, daß eine Erhöhung der Zahl der durch den Klassifikator zu unterscheidenden Klassen beim Quadratmittelklassifikator die $d_k = \text{const}$ Kurven von allen Klassen vollständig verändert. Beim abstandsmessenden Klassifikator und Normalverteilungsklassifikator muß zur Bestimmung des Ortes $d_k = \text{const}$ ausschließlich die klassenspezifische Kovarianzmatrix K_k bekannt sein, die sich aus den Stichprobenelementen einer Klasse berechnen läßt. Beim Quadratmittelklassifikator ist die Matrix G_k der Klasse k von allen anderen Kernmatrizen G_j abhängig. Sie kann somit nicht getrennt berechnet werden.

Die Darstellungen der Ortskurven $d_k = \text{const}$ werden damit sicherlich etwas übersichtlicher, wenn zur Berechnung der Koeffizientenmatrix des Quadratmittelklassifikators tatsächlich nur diejenigen Zeichenklassen herangezogen werden, die

auch in den Darstellungen durch Repräsentanten gekennzeichnet sind, d. h. , die Klassen "1", "2", "0". In Bild 59 sind diese Ortskurven in der Ebene $w_1 w_2$ gezeichnet. Für die Kennung "2" und "1" ergeben sich Hyperbeln, für die Kennung "0" Ellipsen, deren Mittelpunkt weit außerhalb des Zeichenbereichs liegt. An diesem Beispiel kann demonstriert werden, daß die Kegelschnitte $d_k = 0$ für alle drei Klassen stets zentral durch die fremden Klassengebiete verlaufen, während die Kegelschnitte $d_k = 1$ ausschließlich durch die eigenen Klassengebiete verlaufen.

Diese Beispiele wurden ausgewählt, um das prinzipielle Verhalten des Quadratmittelklassifikators hinsichtlich der erzeugten Orte konstanter Schätzung deutlich zu machen. Die starke Einschränkung in Bezug auf die Dimension des Merkmalsraumes und der Klassenzahl war zu dieser Demonstration notwendig.

Abschließend wird noch an dem zuletzt diskutierten Beispiel von drei Klassen gezeigt, wie die Zeichen, deren Lage im zweidimensionalen Merkmalsraum Bild 59 wiedergibt, in den Entscheidungsraum durch die Vorschrift $d = A^T x(w)$ abgebildet werden. In Bild 60 ist die Verteilung der Punkte im Entscheidungsraum dargestellt. Bei einem Drei-Klassen-Problem bilden sich diese Punkte auf einer zweidimensionalen Ebene ab, die durch die Spitzen der drei Basisvektoren

y_1, y_2, y_3 , d. h., durch die Punkte $(1, 0, 0)$, $(0, 1, 0)$, $(0, 0, 1)$ im dreidimensionalen Entscheidungsraum verläuft. In diese Ebene wird die Zeichenebene gelegt. Damit ergibt sich als Ziel der Schätzung, daß sämtliche Punkte entsprechend der Klasse auf einer Ecke des gezeichneten gleichseitigen Dreiecks liegen sollen. Der Fehlervektor f ist dann 0 . Wie aus Bild 60 hervorgeht, sammeln sich die Punkte um diese Ecken.

Um deutlich zu machen, welchen Einfluß die Zahl der verwendeten Koeffizienten der Schätzgleichung auf die Verteilung

der Punkte im Entscheidungsraum hat, wird zur Schätzung d_k aus Demonstrationsgründen ein in den beiden Basismerkmalen $w_1 w_2$ vollständig kubisches Polynom verwendet. Das Ergebnis zeigt Bild 61. Sämtliche Punkte häufen sich wesentlich konzentrierter an den Ecken des Dreiecks. Der mittlere Betrag des Fehlervektors f , der bei fehlerloser Klassifizierung die Reststreuung S^2 ist, wurde erheblich verkleinert.

5. Leistungsvergleich der betrachteten quadratischen Klassifikatoren am Beispiel der Handblockschriftstichprobe

In Kapitel 4.3. wurden die betrachteten Klassifikatoren, deren Schätzfunktionen einheitlich vom Typus des quadratischen Polynoms sind, deren Optimierungskriterien zur Einstellung der Koeffizienten jedoch völlig verschieden sind, einander gegenüber gestellt. Wesentlich dabei war, Gemeinsamkeiten und Unterschiede, die aufgrund des jeweiligen Ansatzes prinzipiell vorliegen, aufzuzeigen, und so zum Verständnis der Konzepte beizutragen. Anschauliche Deutungen und Interpretationen im darstellbaren zweidimensionalen Merkmalsraum sollten dies unterstützen.

Im Hinblick auf den Einsatz von Erkennungssystemen interessiert jedoch eigentlich mehr, welches Konzept für praktische Aufgaben leistungsfähiger ist. Um dies festzustellen, müssen Messungen an den verschiedenen Systemen unter realistischen Bedingungen durchgeführt werden. Dies bedeutet, zur Feststellung der Leistungsfähigkeit realistisches, umfangreiches Stichprobenmaterial heranzuziehen. Wir werden hierzu die Handblockschriftstichprobe HO verwenden. Diese Stichprobe wurde für alle Experimente sowohl zum Lernen wie zum Testen verwendet. Bei umfangreichen Stichproben ist dies zulässig. Es ergeben sich dabei dieselben Resultate, wie wenn die gesamte Stichprobe in Test- und Lernstichprobe aufgeteilt wird. Ausführliche Diskussionen zu den Stichworten "Stichprobenumfang" und "Test- und Lernstichprobe" finden sich in

/L2, K2, K7/.

Nun sind aussagefähige Vergleiche nur dann möglich, wenn bei gleichem investierten Aufwand die erreichten Leistungsmaße der jeweiligen Systeme einander gegenübergestellt werden. Es ist sinnvoll, den Aufwand, der im praktischen Betrieb des Erkennungssystems erforderlich ist, mit der pro Klasse abzuspeichernden Koeffizientenzahl zu messen. Diese Zahl der benötigten Koeffizienten entspricht gleichzeitig der Länge des Schätzpolynoms. Darüberhinaus sollte der Adaptionaufwand der zur optimalen Einstellung der Koeffizienten des Schätzpolynoms aufgebracht werden muß, getrennt betrachtet und gewertet werden. Dieser Aufwand ist für jedes Erkennungssystem nur einmal notwendig, das dann in beliebig vielen Lesemaschinen eingesetzt werden kann.

Als Leistungsmaße der verschiedenen Systeme dienen die gemessenen Fehler- und Rückweisungsrate. Nun können aber diese Raten durch Einstellen einer internen Zurückweisungsschwelle verändert werden. Dabei ist es stets möglich, die Fehlerzahl eines Systems auf Kosten der Zahl der zurückgewiesenen Zeichen zu verringern. Bei praktisch eingesetzten Systemen wird dabei der Arbeitspunkt so gewählt, daß eine gerade noch tolerierbare Fehlerrate nicht überschritten wird. Die Größe dieser Fehlerrate ist natürlich stark vom jeweiligen Anwendungsfall abhängig. So können beispielsweise bei zusätzlich vorgesehenen Kontextprüfungen (Prüfzeichenkontrolle) wesentlich höhere Fehlerraten in Kauf genommen werden als bei Systemen ohne zusätzliche Prüfungen. Daher interessiert nun, wie groß bei eingestellter Zurückweisungsschwelle die Fehlerrate und die Rate der zurückgewiesenen und damit nicht ausgewerteten Zeichen ist, d. h., zu welchem Preis die Fehlerrate erkaufte worden ist.

Um den Arbeitspunkt eines Systems durch die Wahl einer internen Zurückweisungsschwelle einstellen zu können, muß zu-

nächst ein geeignetes Zurückweiskriterium gewählt werden. Hierzu eignet sich für alle Systeme die Differenz zwischen bester und zweitbesten Schätzung der Rückschlußwahrscheinlichkeit

$$DIF = d_{k(\max 1)} - d_{k(\max 2)}$$

Diese Größe bringt zum Ausdruck, wie sicher die Entscheidung des Systems ist. Bei klassifizierten Zeichen mit sehr kleinen DIF-Werten ist die Wahrscheinlichkeit einer falschen Erkennung wesentlich größer als bei solchen mit größerem DIF-Wert. Die Fehlerrate läßt sich dann verkleinern, wenn die Zeichen, deren DIF-Wert kleiner als eine vorgegebene Schwelle ist, zurückgewiesen werden.

Bei dem Quadratmittelklassifikator bietet sich als Zurückweiskriterium der Betrag des Fehlervektors $\mathbf{f}$ an, der sich aus dem Schätzvektor $\mathbf{d}$ und dem Zielvektor $\mathbf{y}_{ent}$ derjenigen Klasse ergibt, für die sich das Erkennungssystem entschieden hat:

$$\text{Fehlervektor } \mathbf{f} = \mathbf{d} - \mathbf{y}_{ent} \quad (115)$$

$$\begin{aligned} \text{Rückweiskriterium RAD} &= |\mathbf{f}| \quad (116) \\ &= \sqrt{(\mathbf{d} - \mathbf{y}_{ent})^2} \end{aligned}$$

Im Entscheidungsraum ist demnach das Gebiet, in dem Zeichen einer Klasse nicht zurückgewiesen werden, begrenzt durch eine Hyperkugel, deren Mittelpunkt die Spitze des Vektors $\mathbf{y}_{ent}$ und deren Radius die Zurückweiskriteriumsschwelle RAD ist. Zwei zu beliebigen Klassen i und j gehörige Hyperkugeln berühren sich dann, wenn $RAD = \frac{1}{2} \sqrt{2}$ gewählt wird. Der Berührungspunkt liegt auf der Geraden, auf der die Endpunkte von $\mathbf{y}_i$ und $\mathbf{y}_j$ liegen. Bei größeren Werten von RAD durchdringen sich die Hyperkugeln, bei kleineren entstehen große, nicht umschlossene Gebiete. Demnach wird deutlich,

daß eine natürliche Grenze für die Wahl der Zurückweisungsschwelle RAD bei dem Wert $\text{RAD} = \frac{1}{2} \sqrt{2} \approx 0.7$ gegeben ist. Da bei der Schätzung der Rückschlußwahrscheinlichkeiten durch die Polynome $d_k = \mathbf{a}_k^T \mathbf{x}(\mathbf{w})$, deren Koeffizienten im quadratischen Mittel adaptiert wurden, die Bedingung "Summe der geschätzten Wahrscheinlichkeiten ist 1"

$$\mathbf{e}^T \mathbf{d} = 1 \quad (117)$$

($\mathbf{e} = (1, 1, \dots, 1)^T$) nach Gleichung (113b)

erfüllt ist, liegen die Spitzen der Entscheidungsvektoren $\mathbf{d}$ auf der durch Gleichung (117) definierten Hyperebene. Damit ist die Dimension der Hyperkugel, innerhalb derer Zeichen akzeptiert werden $K - 1$, wobei K die Zahl der zu unterscheidenden Klassen ist.

Im Gegensatz zu dem DIF-Kriterium bezieht das RAD-Kriterium sämtliche K Schätzwerte mit ein. Dies trägt dazu bei, daß dieses Zurückweisungskriterium besonders dazu geeignet ist, unsicher klassifizierte Zeichen von den übrigen zu unterscheiden.

Bei der Auswertung der zur Beurteilung der verschiedenen Systeme herangezogenen Fehlerraten ξ und Zurückweisungsraten ϑ , die über der Zurückweisungsschwelle abgetragen wurden, muß berücksichtigt werden, daß die Fehlerkurven wesentlich weniger statistisch zuverlässig sind als die Zurückweisungskurven. Dies ist unmittelbar Folge des begrenzten Stichprobenumfanges. Bei sehr kleinen Werten von ξ muß diese Rate aus wenigen festgestellten Ereignissen ermittelt werden: klar ist, daß eine Fehlerrate von $\xi < 10^{-4}$ wie sie von Erkennungssystemen tatsächlich erreicht wird, nicht mit großer statistischer Zuverlässigkeit durch Klassifizierung der etwa 20 000 Zeichen der Stichprobe H_0 gemessen werden kann.

Dieses Problem des Stichprobenumfangs ist aber hier nicht so bedeutend, da keine übertragbaren Aussagen über Fehler-raten gemacht werden. Die Verwendung anderer Stichproben ergibt mit Sicherheit andere Ergebnisse. In dieser Arbeit wird jedoch nur ein Vergleich verschiedener Systeme an genau dem gleichen Stichprobenmaterial durchgeführt.

5.1. Leistungsmessung an den Klassifikatoren nach 4.1.

In Bild 62 wurden die Fehlerraten, die sich ohne Rückwei-sung von Zeichen ergaben, abgetragen über der Anzahl M der Merkmale w_i , die zur Bestimmung der Schätzung d_k (Gleichung (81) und (83a)) benutzt wurden. Gemessen wurden dabei die Werte bei $M = 10, 20, 30, 40$. Aus den Kurven geht deutlich hervor, daß der Normalverteilungsklassifikator bei jedem M eine geringere Fehlerrate aufweist, als der abstandsmessen-de Klassifikator. Beide Kurven verlaufen dabei über den gan-zen Wertebereich von M mit nahezu konstantem Abstand. Wäh-rend es für kleine M noch möglich ist, durch zusätzliche Berücksichtigung von Merkmalen eine erhebliche Reduzierung der Fehlerrate zu erreichen, so ist dies ab etwa $M = 25$ Merkmalen kaum mehr möglich, d. h., die Kurven gehen deut-lich in die Sättigung. Die Erhöhung der Zahl der Merkmale von $M = 30$ auf $M = 40$ reduziert beispielsweise die Fehler-rate bei beiden Systemen nur noch um 0.7%.

Diese Ergebnisse können wie folgt interpretiert werden. Wird der $N \approx 200$ -dimensionale Zufallsprozess $\{w/k\}$ in sehr wenigen Dimensionen beschrieben ($M \approx 10$), die Vertei-lung der Merkmale in diesem Raum durch eine Normalvertei-lung modelliert, und die sich daraus ergebende Schätzung d_k der Rückschlußwahrscheinlichkeit als Unterscheidungsfunktion in einem Erkennungssystem verwendet, so ergeben sich gute Erkennungsraten (etwa 97.5%). Wesentlich bessere Erkennungs-ergebnisse lassen sich bei einer derartig drastischen Re-duktion der Dimension des Merkmalsraums kaum erwarten.

Begnügt man sich mit dieser Erkennungsleistung jedoch nicht, sondern versucht durch Verwendung weiterer Merkmale eine vollständigere Beschreibung des Zufallsprozesses durch das M-dimensionale Modell zu erhalten, so zeigt sich bereits im Bereich von 30 bis 40 Dimensionen keine Leistungssteigerung des Klassifikators mehr. Dies deutet darauf hin, daß der Prozess $\{w/k\}$ durch die Normalverteilung nicht genügend genau modellierbar ist, insbesondere an den Stellen im Merkmalsraum, die sehr dünn von Zeichen besetzt sind. Diese wenigen Zeichen werden dabei in erhöhtem Maße falsch klassifiziert. Diese Ergebnisse stimmen überein mit den Schlußfolgerungen, die aus den gemessenen Verteilungen des Abstandsmaßes Q_k und dem Modell χ^2 -Verteilung gezogen wurden. In 40 Dimensionen war die Abweichung von Modell und gemessener Kurve besonders an Stellen außerhalb des Dichtezentrums größer als in 10 Dimensionen.

Die unterschiedlichen Klassifizierungsergebnisse von Normalverteilungsklassifikator und abstandsmessendem Klassifikator lassen sich nur durch große Unterschiede in dem Wert der klassenspezifischen Determinanten der Kovarianzmatrix K_k erklären, d. h., die benutzte statistische Annahme stimmt nicht. Zur Demonstration wurden die Werte hierzu für alle Klassen und verschiedene Werte M in Bild 63 tabelliert. Diese stark differierenden Werte lassen auf sehr unterschiedliche Größen der Klassegebiete im M-dimensionalen Raum schließen. Für den 2-dimensionalen Merkmalsraum konnte dies bereits festgestellt werden. Wie in Bild 40 und 41 zu sehen ist, liegen die Zeichen der Kennung "1" wesentlich gedrängter zusammen als die Zeichen der Kennung "0".

In Bild 64 wurde für den Normalverteilungsklassifikator Rückweisungsrate β und Fehlerrate ε über dem Wert der Zurückweisungsschwelle DIF abgetragen. Man kann daraus ablesen, daß eine Fehlerrate von etwa $\varepsilon = 2\%$ erreicht wird, wenn bei 40 verwendeten Merkmalen ungefähr 7%, bei 30 ver-

wendeten Merkmalen etwa 1,5% und bei 20 verwendeten Merkmalen 5% der Zeichen zurückgewiesen werden. Fehlerfreiheit erhält man für $M = 40$ bei einer Zurückweisungsrate von etwa $\beta = 2\%$. An dieser Stelle sei noch einmal auf die gegenüber den β Kurven größeren statistischen Vertrauensintervalle der ε Kurven hingewiesen.

Die Vorteile von Erkennungssystemen, die Annahmen über den Verteilungstyp der Dichte der Merkmale benutzen, liegen im wesentlichen in dem geringen Adaptionsaufwand. Ein Maß dafür ist die Rechenzeit zur Ermittlung der Parameter K_k^{-1} und det K_k , die von einem Großrechner benötigt wird. In Bild 65 sind die entsprechenden Rechenzeiten (TR440) über der Zahl der Merkmale abgetragen. Sie betragen für alle Systeme nur wenige Minuten. Diese Zeiten sind für Normalverteilungsklassifikator und abstandsmessenden Klassifikator in etwa gleich. Um den im praktischen Betrieb erforderlichen Realisierungsaufwand, der mit der Zahl L der pro Klasse abzuspeichernden Koeffizienten gemessen wird ($\hat{n}$ Länge des Schätzpolynoms) der jeweils erreichten Leistung gegenüberzustellen, wurde in Bild 66 die absolute Fehlerzahl bei erzwungener Klassifikation ($\beta = 0$) abgetragen über L . Auch hier wird deutlich, daß eine Erhöhung der Koeffizientenzahl ab $L = 200$ keine wesentliche Reduzierung der Fehlerzahl mehr ermöglicht.

5.2. Leistungsmessung an dem Klassifikator nach 4.2.

Bei dem Lösen des linearen Gleichungssystemes nach Gleichung (96) durch den schrittweisen Regressionsprozess ist in jedem Bearbeitungszustand der Momentenmatrix die aktuell erreichte Reststreuung S^2 (Gleichung (85) mit Berücksichtigung von Gleichung (92)) verfügbar. Der Verlauf der Reststreuung S^2 kann damit über der Anzahl L der zur Schätzung benutzten Variablen - L ist dabei auch die Polynomlänge und gleichzeitig die Zahl der pro Klasse verwendeten Koeffizien-

ten - abgetragen werden. Berechnet wurden verschiedene Klassifikatoren mit vollständig quadratischem Ansatzpolynom. Die Klassifikatoren unterscheiden sich nur in der Zahl M der verwendeten Merkmale. Die folgende Tabelle gibt die Zuordnung von Klassifikatornamen, Länge L_{VQ} des Ansatzpolynoms und verwendeter Zahl M von Merkmalen wieder.

M	L_{VQ}	Klassifikatorname
10	66	HO1HTQ
20	231	HO2HTQ
30	496	HO3HTQ
40	861	HO4HTQ

Bei den Bezeichnungen der Klassifikatoren steht HO für "Stichprobe HO", die folgende Zahl X ($1 \leq X \leq 4$) im Namen weist auf die Zahl M der verwendeten Merkmale ($M = 10X$), die Abkürzung HTQ steht für "Hauptachsentransformation mit quadratischem Klassifikator".

Für den interessierenden Bereich ($S^2 < 0,13$) ist in Bild 67 die Reststreuung über der Länge L des Schätzpolynoms für drei der bezeichneten Klassifikatoren abgetragen. Entsprechend der Länge ihres Ansatzpolynoms L_{VQ} enden diese Kurven bei den Werten 231 ($M = 20$), 496 ($M = 30$), 861 ($M = 40$). Aus den Kurven ist abzulesen, daß der erreichte Endwert der Reststreuung bei HO2HTQ etwa doppelt so groß ist, wie bei HO4HTQ. Es zeigt sich, daß dabei jeweils derjenige Klassifikator bei gleicher Schätzgleichungslänge eine kleinere Reststreuung S^2 besitzt, der mehr Merkmale w_i im Ansatzpolynom verwendet.

In Bild 68 sind die durch Simulation des Erkennungssystems bestimmten Zurückweisungs- und Fehlerraten über der Zurückweisungsschwelle, nämlich dem Betrag RAD des Fehlervektors f abgetragen. Dabei wurde für alle Klassifikatoren das in den Merkmalen w_i vollständig quadratische Schätzpolynom verwendet.

Aus dem Diagramm ist abzulesen, daß eine Erhöhung der Zahl M der verwendeten Merkmale w_i von 10 auf 20 eine deutliche Verringerung sowohl der Fehlerrate ξ wie der Rückweisungsrate ζ ergibt. Mit zunehmendem M wird die Steigerung der Leistung etwas geringer, jedoch ist ein Sättigungseffekt, d. h., ein Stagnieren der Leistung trotz Erhöhung der Schätzgleichungslänge L nicht erkennbar.

Bild 69 zeigt die Fehlerrate ξ über der Zahl M der Merkmale w_i . Die obere Kurve ist dabei die Fehlerkurve bei erzwungener Klassifizierung, die untere diejenige, die erreicht wird, wenn die Zurückweisungsschwelle auf $RAD = 0.7$ eingestellt wird. Gemessen wurde wiederum an den Stellen $M = 10, 20, 30, 40$. Das Schätzpolynom ist in allen Fällen vollständig quadratisch. Man erkennt, in welchem Maße die Fehlerrate bei dieser Einstellung der Zurückweisungsschwelle gegenüber der erzwungenen Klassifizierung ($\zeta = 0$) verringert wird.

Die zugehörige Rückweisungsrate ζ bei der Schwelle $RAD = 0.7$, die wieder über der Zahl M der Merkmale abgetragen ist, zeigt Bild 70. Auch hier ist das Schätzpolynom wieder vollständig quadratisch. Man erkennt mit zunehmendem M eine deutliche Abnahme der Rückweisungen.

Die Tendenz, bei Steigerung der Zahl M der Merkmale weniger Zeichen zurückweisen zu müssen, ist in der halblogarithmischen Darstellung in Bild 71 besser sichtbar. In diesem Diagramm wurde die Zurückweisungsrate ζ , die notwendig

ist, um ein vorgegebenes ξ ($10^{-3}, 3 \cdot 10^{-4}$) zu erhalten, über M abgetragen. Man erkennt, daß alle Kurven in dieser Darstellung gut durch Geraden zu approximieren sind. Dabei kann abgelesen werden, daß bei gleicher vorgegebener Fehlertrate die Erhöhung der Zahl der verwendeten Merkmale w_1 um 10 eine Verringerung der Zurückweisungsrate um etwa den Faktor 5 ergibt.

Aus den bisher diskutierten Diagrammen geht deutlich hervor, daß bei den Experimenten mit der Stichprobe HO ein Stagnieren der Leistung des Quadratmittelklassifikators, auch bei großer Zahl M von Merkmalen, - im Gegensatz zu dem Normalverteilungsklassifikator und abstandsmessenden Klassifikator - nicht festzustellen ist.

Am Beispiel des Quadratmittelklassifikators $HO4HTQ$, dessen Ansatzpolynom in $M = 40$ Merkmalen vollständig quadratisch ist, soll nun die Leistungsfähigkeit von Klassifikatoren demonstriert werden, die bei dem Regressionsprozess (mit MS-Strategie-Steuerung) als Zwischenresultate anfallen. Die Schätzgleichungslänge L dieser Klassifikatoren kann im Bereich $L < L_{VQ}$ beliebig eingestellt werden. In Bild 72 sind die Zurückweisungs- und Fehlerkurven über dem Betrag des Fehlervektors RAD abgetragen. Die Länge der Schätzgleichung wird dabei beginnend bei $L = 100$, jeweils um 100 erhöht. Der maximale Wert für L ergibt sich bei den 40 Merkmalen zu $L = L_{VQ} = 861$. Die η und ξ Kurve für diese Schätzgleichungslänge wurde bereits in Bild 68 gezeichnet.

An den Diagrammen ist die Leistungszunahme der Systeme mit zunehmender Schätzgleichungslänge durch deutliches Abnehmen sowohl von Fehler- wie Rückweisungsraten erkennbar. Vergleicht man diese Kurven mit den η und ξ Diagrammen von Bild 68, die für vollständig quadratische Schätzpolynome gemessen wurden, so zeigt sich, daß bei gleicher Schätzgleichungslänge L die Kurven in Bild 72 ein wenig günstiger

liegen. Dies bedeutet, daß ein Erkennungssystem mit Schätzpolynomen, die zur Messung der Kurven in Bild 72 verwendet wurden, etwas weniger Zeichen zurückweist bei gleichzeitiger Verkleinerung der Fehlerrate. Man kann dies mit der Art und Weise erklären, mit der Struktur und Koeffizienten der Schätzpolynome des Klassifikators H04HTQ berechnet werden. Hier stehen zu Beginn des Regressionsprozesses eine größere Menge von Merkmalen w_i und Merkmalsverknüpfungen $w_i w_j$ als potentielle Kandidaten zum Einbau in die Schätzgleichung zur Verfügung, von denen nun schrittweise diejenigen herangezogen werden, die im Sinne der Forderung nach minimaler Reststreuung die besten sind. Die Terme w_i , $w_i w_j$, die im vollständig quadratischen Polynom mit $M < 40$ Merkmalen enthalten sind, sind vollständig darin enthalten. Geht man davon aus, daß nun eine Strategie zur Auswahl der Variablen der Schätzgleichung gewählt wird, die bei vorgegebenem L und fester Struktur des Ansatzpolynoms, eine kleinstmögliche Reststreuung S^2 bei jeweils optimierten Koeffizienten ermöglicht, so kann gefolgert werden, daß die Kurven in Bild 72 nicht unter den zugehörigen Kurven von Bild 68 liegen können.

Aus der Tatsache, daß diese Kurven jedoch nur einen kleinen Abstand zu den Kurven gleicher Schätzgleichungslänge L in Bild 68 aufweisen, kann man den Schluß ziehen, daß die Merkmale w_i mit niedriger Rangnummer (wichtige Merkmale) als Variable im Schätzpolynom sehr effektiv sind.

Wird während der Matrixbearbeitung die Reihenfolge, in der die zur Verfügung stehenden Variablen x_i in die Schätzgleichung eingebaut werden, ebenfalls festgehalten, so kann festgestellt werden, ob die Rangreihenfolge unter den Merkmalen w_i auch Schlüsse auf die Rangreihenfolge unter den Einflußgrößen des quadratischen Klassifikators, d. h., unter den Merkmalsverknüpfungen zuläßt. Wir betrachten hierzu die Rangreihenfolge unter den 860 Variablen w_i , $w_i w_j$ des

Klassifikators HO^4HTQ und verwenden dafür die Darstellung von Bild 73 und 74. Jeder der 860 Variablen, die Merkmalsquadrate w_i^2 , Merkmalsverknüpfungen $w_i w_k$ oder die Merkmale w_i selbst sind, ist in dieser Darstellung ein Platz in einer Dreiecksmatrix zugewiesen. Die erste Spalte (Index $k = 0$) gehört dabei zu den linearen Gliedern des Ansatzes, den Merkmalen w_i selbst. Man kann nun an die, jedem Element des Ansatzes zugeordnete Stelle die entsprechende Rangnummer eintragen. Dies ist in Bild 73 durchgeführt worden. Man kann aber auch, um eine bessere Übersicht zu erhalten, eine quantisierte Form der Darstellung wählen: Rangnummern unter 200 werden durch ein Kreuz (x), Rangnummern zwischen 201 und 400 durch einen Kreis O gekennzeichnet (Bild 74).

Aus diesen Darstellungen ist nun abzulesen, daß Merkmale mit hoher Stellung in der Rangliste, d. h., niedrigen Indizes i, j auch als Produktterme $w_i w_j$ für sehr wichtig erachtet werden. Dieses Ergebnis ist keineswegs selbstverständlich, denn Untersuchungen an linearen und quadratischen Klassifikatoren mit direkter Verwendung der Merkmale $v_i / F_4 /$ haben gezeigt, daß wichtige Merkmalsverknüpfungen oft aus Merkmalen gebildet werden, die als lineare Elemente bei dem Regressionsprozess eliminiert, d. h., für bedeutungslos erachtet werden. Dies ist hier offensichtlich nicht der Fall, was rein optisch schon an der sehr starken Häufung der Kreuze für kleine i und k in Bild 74 zu erkennen ist. Daraus ergibt sich nun die Folgerung, daß es nicht nötig ist, schon im Ansatzpolynom sehr viel mehr Variable zur Verfügung zu stellen, um bei Blockierung der Schätzungsgleichungslänge eine bestmögliche Auswahl zu garantieren. Werden in dem hier betrachteten Beispiel aus den 860 angebotenen nur 400 Variable zur Schätzung herangezogen, so hätte man, um fast die gleiche Auswahl zu erhalten, lediglich den vollständig quadratischen Klassifikator mit etwa 33 Merkmalen anzubieten brauchen, die restlichen Variablen werden ohnehin auch als Verknüpfungen kaum zur

Schätzung herangezogen. Wichtigstes Ergebnis hierbei ist, daß Merkmale w_i , die durch die HAT produziert wurden, sowohl als solche, als auch als Verknüpfungen bedeutend für die Schätzung der Klassenzugehörigkeit sind.

Um den für den Quadratmittelklassifikator notwendigen Adaptionsaufwand aufzuzeigen, der im wesentlichen in der Aufstellung und dem Lösen des linearen Gleichungssystems (Gleichung (96)) besteht, wurde in Bild 75 die zur Bestimmung der Koeffizientenmatrix A notwendige Rechenzeit (wiederum TR 440 Rechnerkernzeit) über der Zahl M der verwendeten Merkmale w_i abgetragen. Man erkennt eine sehr starke Zunahme der Rechenzeit bei Erhöhung von M . Dieser relativ hohe Adaptionsaufwand ist für Klassifikatoren typisch, die keine statistischen Hypothesen über den zeichenerzeugenden Zufallsprozess $\{w, k\}$ benutzen. Die statistischen Eigenschaften dieses Prozesses werden erst in der Adaptionsphase bestimmt. Dieses Diagramm ist unmittelbar mit Bild 65 zu vergleichen, in dem der Adaptionsaufwand für Normalverteilungsklassifikator und abstandsmessenden Klassifikator dargestellt ist.

Entsprechend Bild 66 wurde in Bild 76 die Fehlerzahl bei erzwungener Klassifikation über der Länge L des Schätzpolynoms abgetragen. Gemessen wurde dabei an den Stellen $L = 66, 231, 496, 861$, und zwar für Klassifikatoren mit vollständig quadratischem Schätzpolynom (H01HTQ, H02HTQ, H03HTQ, H04HTQ). Man erkennt eine starke Abnahme der Fehlerzahl bei steigendem L . Ein Sättigungseffekt, d. h., ein Stagnieren der Fehlerzahl trotz Erhöhung von L ist hier nicht erkennbar.

Der Quadratmittellansatz bietet die Möglichkeit, Erkennungssysteme sowohl mit vollständig wie mit unvollständig quadratischen Schätzfunktionen zu realisieren. Wird ein Erkennungssystem mit dem nicht vollständig quadratischen Poly-

nom realisiert, so ist es sinnvoll, den gesamten Aufwand für Klassifikation und Transformation $V \rightarrow W$ zu betrachten, und diesem den Aufwand für das System mit vollständig quadratischem Polynom gegenüberzustellen. Da im Falle des unvollständig quadratischen Polynoms die Berechnung von mehr Merkmalen w_i nötig ist, muß dieser zusätzliche Aufwand mit der Leistungssteigerung in Beziehung gebracht werden. Hierzu führen wir als Maß für den Aufwand eines Erkennungssystems die Zahl Z der abzuspeichernden Koeffizienten für Transformation und Klassifikation ein, und als Maß für die Leistung die Zurückweisungsrate $\mathcal{S}$ bei der Zurückweisungsschwelle $RAD = 0.7$. Bild 77 gibt den Verlauf von $\mathcal{S}_{RAD=0.7} = f(Z)$ für die drei Klassifikatoren $HO2HTQ$, $HO3HTQ$, $HO4HTQ$ wieder. Die Koeffizientenzahl wird dabei ausschließlich durch die Schätzgleichungslänge L verändert und ist wie folgt definiert:

$$Z = Z(L) = M \cdot (N+1) + K \cdot L \quad (118)$$

- M Zahl der verwendeten Merkmale w_i
- N Zahl der Rasterbildelemente (216)
- K Klassenzahl
- L Schätzgleichungslänge

Alle Kurven enden dann, wenn zur Schätzung der Klassenzugehörigkeit alle zur Verfügung stehenden Variablen verwendet werden, d. h., das Schätzpolynom vollständig quadratisch ist. Die Grenzkurve, die zu den Systemen mit vollständig quadratischen Schätzpolynomen gehört, ist ebenfalls in das Diagramm eingezeichnet. Hier wird die Koeffizientenzahl durch die Zahl der Merkmale M verändert. Entsprechend Gleichung (95) ergibt sich die Schätzgleichungslänge L , und damit die Koeffizientenzahl Z zu

$$Z = Z(M) = Z(L(M)).$$

Im Gegensatz zu den Kurven $Z(L)$ nach Gleichung (118) kann sich hier durch Vergrößern von M keine zusammenhängende Folge für Z bilden. Die gezeichnete Kurve ergibt sich durch Verbinden der entsprechenden Punkte. Aus dem Diagramm geht hervor, daß für jede Koeffizientenzahl Z ein System mit vollständig quadratischem Schätzpolynom weniger Rückweisungen erfordert als ein System mit unvollständig quadratischem Schätzpolynom. Der Unterschied der entsprechenden Rückweisungsrate ist dabei umso größer, je mehr die Polynomstruktur von der Struktur des vollständig quadratischen Polynoms abweicht.

In Kapitel 4.2. wurde gezeigt, daß unter der Bedingung von frei wählbarem Funktionstyp $d_k(\mathbf{W})$ sich bei der Minimierung der Reststreuung S^2 die Regressionsfunktion ergibt. S^2 wird dann 0, wenn stets mit der Wahrscheinlichkeit $p(k/\mathbf{W}) = 100\%$ aus dem Merkmalsvektor $\mathbf{W}$ auf die Klasse k geschlossen werden kann. Nun ergibt sich bei realistischen Erkennungssystemen mit Polynomen von zweitem Grad im allgemeinen keine verschwindende Reststreuung. Der Zusammenhang zwischen S^2 und der Fehlerzahl, der aus den Experimenten ermittelt wurde, ist in Bild 78 dargestellt. Hier wird deutlich, daß abnehmende Reststreuung mit erheblicher Abnahme der Fehlerzahl verbunden ist.

5.3. Gesichtspunkte zur praktischen Anwendung von Klassifikatoren nach 4.2.

Im Zusammenhang mit der praktischen Anwendung des Quadratmittelklassifikators sind noch weitere Themen von Interesse. Einmal ist die Beantwortung der Frage wichtig, ob nicht in irgendeiner Weise an Adaptionsaufwand gespart werden kann, der beim Quadratmittelklassifikator doch so deutlich größer ist als beim Normalverteilungsklassifikator und abstandsmessenden Klassifikator. Sicherlich ist dieser Aufwand insbesondere dann störend, wenn sehr wenige Lesemaschinen mit

gleichem Erkennungssystem hergestellt werden. Bei größeren Stückzahlen verringert sich der Anteil der Adaptionskosten an den Gesamtkosten.

Weiter interessiert, welchen Einfluß die Quantisierung der Merkmale w_i auf die Leistung des Erkennungssystem hat. Um die Klassifikationszeit eines Zeichens klein zu halten, sollten einerseits die Quantisierung möglichst grob sein (Verwendung von wenigen Binärstellen), andererseits ist es erwünscht, die dadurch bedingte Leistungseinbuße nicht zu groß werden zu lassen.

Schließlich interessiert noch, ob die Leistungsgrenze des Quadratmittelklassifikators bereits erreicht wurde, oder ob diese Leistung nicht noch verbessert werden kann, etwa indem in einem neuen Adaptionsprozess unsicher erkannte Zeichen verstärkt berücksichtigt werden /B3, S4/.

In diesem Themenkreis werden nun experimentelle Ergebnisse diskutiert. Berechnet wurden drei Klassifikatoren mit vollständig quadratischem Ansatzpolynom. Es wurden dabei jeweils $M = 40$ Merkmale w_i verwendet. Die Klassifikatoren unterscheiden sich nur in der Zahl der zur Quantisierung der Merkmale w_i benutzten Binärstellen. Dabei wurde eine Quantisierung mit gleicher Stufung gewählt, wodurch der Wertebereich $-5 \dots +5$ in $2^X - 3$ Quantisierungsintervalle mit zwei zusätzlichen Randfächern eingeteilt wurde. X ist dabei die Zahl der Binärstellen einschließlich Vorzeichen. Mit diesen quantisierten Merkmalen wurde ein erneuter Adaptionsprozess durchgeführt, der die jeweiligen Koeffizienten der Schätzpolynome im Sinne minimaler Reststreuung einstellt. Die Namen der Klassifikatoren weisen auf die Zahl der verwendeten Binärstellen (hier jedoch ohne Vorzeichen), die zur Quantisierung der w_i verwendet wurden.

NAME	QUANTISIERUNG	X/BIT
HO4HTQ	38 BIT + Vorzeichen (TR440 REAL Darstellung)	39
HO45HQ	5 BIT + Vorzeichen	6
HO44HQ	4 BIT + Vorzeichen	5
HO43HQ	3 BIT + Vorzeichen	4

Der Klassifikator mit REAL Darstellung der Merkmale wurde bereits diskutiert und wird hier nur zu Referenzzwecken erwähnt.

Bild 79 gibt den Verlauf der Reststreuung S^2 über der Länge L des Schätzpolynoms für die 4 tabellierten Klassifikatoren wieder. Man erkennt, daß sich bei einer 6 BIT Quantisierung der Merkmale w_i nahezu der gleiche Verlauf der Reststreuung ergibt wie bei FTN Real Darstellung der w_i . Es zeigt sich weiter, daß sich die Kurve bei einer 4 BIT Quantisierung recht deutlich von den übrigen abhebt.

In Bild 80 sind die Rückweisungs- und Fehlerkurven dieser vier Klassifikatoren über dem Betrag RAD des Fehlervektors f abgetragen. Auch hier zeigt sich deutlich, daß eine 6 BIT Quantisierung der Merkmale w_i die Leistung des Erkennungssystems nur in sehr geringem Maße verringert. Die Erhöhung der Zurückweisungsrate β bei einer 5 BIT Quantisierung ist ebenfalls nicht sehr groß. In beiden Fällen bleibt die Fehlerkurve unverändert. Ein deutlicher Leistungsabfall tritt erst bei einer 4 BIT Quantisierung ein.

Es wurde bereits erwähnt, daß die Adaptionkosten des Quadratmittelklassifikators darin bestehen, das Gleichungs-

system (Gleichung (96)) aufzustellen (d. h., im wesentlichen die Momentenmatrix $E \{ \mathbf{x} \mathbf{x}^T \}$ zu berechnen) und dieses zu lösen. Als Aufwandsmaß hierzu wurde die zur Lösung dieser Aufgabe benötigte Rechnerkernzeit eingeführt. Nun kann durch die Quantisierung der Merkmale w_i erheblich Aufwand bei der ersten Teilaufgabe, der Berechnung der Momentenmatrix gespart werden. Da zur Lösung dieser Teilaufgabe bei dem vorliegenden Quadratmittelklassifikator wesentlich mehr Rechenzeit erforderlich ist, als zur Lösung der zweiten, können damit auch die gesamten Adaptionskosten beträchtlich verringert werden. Die zur Berechnung der Momentenmatrix notwendige Rechnerkernzeit ist in Bild 31 über der Zahl X der zur Quantisierung der Merkmale w_i benutzten BITS dargestellt. Diese Grafik zeigt, wie stark die Rechenzeit mit abnehmendem X abfällt. Mit eingezeichnet als Balken im 4. Quadranten des Koordinatensystems ist die jeweils zugehörige Fehlerzahl bei erzwungener Klassifizierung. Man sieht, daß die Fehlerzahl erst bei der 4 BIT Quantisierung stark zunimmt. Eine 6 BIT Quantisierung ermöglicht die gewünschte Reduktion der Rechenzeit und führt dabei zu keiner Zunahme der Fehlerzahl.

Um diese Leistung des Erkennungssystems zu erhöhen, ohne dafür mehr Koeffizienten für die Schätzfunktion bereitstellen zu müssen, nützen wir die Tatsache aus, daß reale Systeme mit fest eingestellten Rückweisungsschwellen, die sinnvoll gewählt wurden, arbeiten (etwa bei $RAD \approx 0.7$). Das Ziel ist demnach, Zurückweisungsrate ξ und Fehlerrate ξ an diesem Arbeitspunkt möglichst klein zu halten. Man muß daher versuchen, die Koeffizienten des Schätzpolynoms so einzustellen, daß die wenigen unsicher erkannten Zeichen, unter denen sich die falsch klassifizierten Zeichen befinden, mit ausreichender Sicherheit erkannt werden. Dies kann auf Kosten der Erkennungssicherheit der gut erkannten Zeichen geschehen. Allerdings muß dabei vorsichtig vorgegangen werden, da es unbedingt zu vermeiden ist, daß eine größere Zahl von zuerst sehr sicher erkannten Zeichen

nun plötzlich in den Bereich der unsicheren Kandidaten gelangt.

Diese Neueinstellung der Koeffizienten erfolgt in einer zweiten Adaptionsphase. Die durch den Ausgangsklassifikator schlecht erkannten Zeichen werden abgesondert und zu einer Problemstichprobe zusammengefaßt. Die ursprüngliche Stichprobe wird nun verfärbt, indem die unsicher erkannten Zeichen verstärkt berücksichtigt werden. Nun wird ein zweiter, iterierter Quadratmittelklassifikator auf der Basis der neuen Stichprobe berechnet. Da die Momentenmatrix M der ursprünglichen Stichprobe bereits berechnet wurde, ist eine vollständig neue Berechnung der Momentenmatrix M_{NEU} der verfärbten Stichprobe nicht mehr erforderlich. Es genügt, die Momentenmatrix M_P der abgesonderten Problemzeichen, die aus wesentlich weniger Stichprobenelementen besteht, zu berechnen. Die Matrix M_{NEU} ergibt sich dann zu

$$M_{NEU} = M(1 - \delta) + \delta M_P \quad (98)$$

Dabei gibt δ an, wie stark die Problemzeichen zugemischt werden. In den Experimenten wurde $\delta = 0.2$ gewählt. Dieser Wert hat sich aus einer Reihe von Experimenten als günstig erwiesen /B3, S4/.

In Bild 82 ist die Reststreuung S^2 des Ausgangsklassifikators H04HTQ sowie des iterierten Klassifikators H04HTQ1 über der Länge L des Schätzpolynoms abgetragen. Man sieht, daß die Reststreuungskurve des iterierten Klassifikators deutlich oberhalb derjenigen des Ausgangsklassifikators liegt. Erklärbar ist dieses Verhalten mit der veränderten Situation, die der iterierte Klassifikator vorfindet: durch die verstärkte Berücksichtigung schlecht erkennbarer Zeichen in der neu zusammengestellten Lernstichprobe ergibt sich für den Klassifikator auch ein schwierigeres Erkennungsproblem. Dies äußert sich in einer größeren Reststreuung.

Nach der durchgeführten Adaption an die verfärbte Lernstichprobe wird die Leistung des Klassifikators jedoch an der ursprünglichen, nicht verfärbten Stichprobe gemessen. Dies bedeutet, daß in diesem Fall Lern- und Teststichprobe nicht mehr, wie bei allen bisher durchgeführten Experimenten, identisch sind. Damit gilt auch insbesondere der in Bild 78 dargestellte Zusammenhang von Reststreuung und Fehlerzahl nicht mehr. In Bild 83 sind die Zurückweisungs- und Fehlerkurven des Klassifikators HO4HTQ1 für verschiedene Schätzgleichungslängen L über dem Betrag RAD des Fehlervektors f gezeichnet. Man erkennt die gegenüber dem Ausgangsklassifikator erheblich gesteigerte Leistung. Bei der Zurückweisungsschwelle $RAD = 0.7$ und $L = 861$ ist eine Fehlerrate von weniger als 0.2% abzulesen. Die Rückweisungsrate liegt in diesem Arbeitspunkt bei 0.4%.

Eine Gegenüberstellung der Leistung von Ausgangsklassifikator und iteriertem Klassifikator zeigt Bild 84. Das Schätzpolynom ist für beide Klassifikatoren vollständig quadratisch in 40 Merkmalen ($L = 861$). Man erkennt daraus, daß bei einer festen Zurückweisungsschwelle von $RAD = 0.7$ sowohl die Fehlerrate wie die Zurückweisungsrate um etwa Faktor 10 verkleinert werden konnte. Die Zurückweisungskurven von Ausgangsklassifikator und iteriertem Klassifikator schneiden sich bei einer Schwelle von etwa $RAD = 0.4$. Für kleinere RAD Werte liegt die ξ -Kurve des iterierten Klassifikators oberhalb der des Ausgangsklassifikators. Dies bedeutet, daß die Zahl sehr sicher erkannter Zeichen, die kleine RAD Werte besitzen, zurückgegangen ist. Gerade dieses Verhalten war erwünscht, werden doch dafür die Mehrzahl der bisher falsch klassifizierten Zeichen sicherer und damit richtig erkannt.

Bild 85 zeigt den Verlauf der absoluten Fehlerzahl bei erzwungener Klassifizierung für den Ausgangsklassifikator HO4HTQ und den iterierten Klassifikator HO4HTQ1, abgetra-

gen über der Länge des Schätzpolynoms. Der Abstand dieser Kurven ist sehr deutlich. Trotzdem konnte die Fehlerzahl beim iterierten Klassifikator nicht auf 0 gesenkt werden. Der Grund dafür ist, daß in der verwendeten Handblock-schrift-Stichprobe H0 kein Aussondern derjenigen Zeichen vorgenommen wurde, die auch der Mensch als eventuell nicht lesbar bezeichnen würde. Das Absondern dieser Zeichen kann prinzipiell nicht^f objektiv durchgeführt werden, da hierzu die menschliche Beurteilung der Zeichenqualität notwendig ist. Die Gefahr dieses Aussondern liegt aber gerade darin, durch zu wohlwollendes Verwerfen von Zeichen die Fehler-raten von Erkennungssystemen mit an sich geringer Leistung künstlich zu steigern. Objektive Vergleiche von Systemen werden daher erschwert, wenn nicht gar unmöglich gemacht.

5.4. Vergleich der Klassifikatoren nach 4.1. und 4.2. hinsichtlich der experimentellen Ergebnisse

Die Leistungsfähigkeit von Normalverteilungsklassifikator und abstandsmessendem Klassifikator einerseits, und Quadratmittelklassifikator andererseits, wurde bisher für die jeweiligen Systemkonzepte getrennt betrachtet und gewertet. Dabei wurde insbesondere aufgezeigt, wie sich die System-leistung durch Erhöhung des investierten Aufwandes steigern läßt. Im folgenden sollen nun die experimentellen Ergebnisse der Konzepte unmittelbar einander gegenübergestellt werden. Da die Aufgabenstellung, nämlich die Klassifizierung all der Zeichen, die die Handblockschriftstichprobe H0 enthält, für alle Systeme vollkommen identisch ist, kann ein objek-tiver Vergleich durchgeführt werden. Dabei geht es nun da-rum, die jeweils erbrachten Leistungen der Systeme bei fest vorgegebenem Aufwand einander gegenüberzustellen. Getrennt betrachtet werden wiederum Klassifikations- und Adaptions-aufwand.

Der Klassifizierungsaufwand wird wiederum mit der Anzahl L

der Koeffizienten gemessen, die verfügbar sein müssen, um aus $\mathbf{W}$ durch das Schätzpolynom eine Entscheidungsgröße d_k zu berechnen (L wurde auch mit Schätzgleichungslänge bezeichnet). Für den Normalverteilungsklassifikator sowie den abstandsmessenden Klassifikator ist L nach Gleichung (82) und Gleichung (84) nur von der Zahl M der verwendeten Merkmale abhängig. Für den Quadratmittelklassifikator kann im Prinzip L beliebig eingestellt werden. Bei der Diskussion von Bild 77 ergab sich jedoch, daß es hinsichtlich des Verhältnisses Leistung/gesamte Koeffizientenzahl ökonomischer ist, stets ein vollständig quadratisches Schätzpolynom zu verwenden, d. h., L ist auch hier nur von der Zahl der Merkmale w_i abhängig. Damit ergibt sich für alle drei betrachteten Systeme, daß die Gesamtzahl der benötigten Koeffizienten zur Klassifikation (=Klassifikationsaufwand) nur von der Zahl M der Merkmale w_i abhängt. M kann damit in den diskutierten Diagrammen unmittelbar als Aufwandsmaß betrachtet werden.

Zum Vergleich der Systeme wird der nicht iterierte Quadratmittelklassifikator herangezogen, da nur dieser gerechterweise dem berechneten abstandsmessenden Klassifikator und Normalverteilungsklassifikator gegenübergestellt werden darf. Eine entsprechende Iteration, wie sie beim Quadratmittelklassifikator durch Verfärbung der Lernstichprobe durchgeführt wurde, ist im Prinzip auch für den abstandsmessenden Klassifikator und Normalverteilungsklassifikator möglich. Allerdings ist die Wirkung hinsichtlich der Leistungssteigerung unbekannt, da entsprechende Experimente nicht durchgeführt wurden.

In Bild 86 ist die absolute Fehlerzahl ohne Rückweisungen über der Zahl M der verwendeten Merkmale als Aufwandsmaß abgetragen, und zwar für alle drei Systeme. In dem Bild erkennt man, daß die zu dem Quadratmittelklassifikator (HO4HTQ) gehörige Kurve die beiden anderen Kurven schnei-

det. Für $M < 17$ Merkmale produziert der abstandsmessende Klassifikator, für $M < 25$ der Normalverteilungsklassifikator weniger Fehler als der Quadratmittelklassifikator. Man erkennt deutlich, daß sich beim Quadratmittelklassifikator durch Aufwandserhöhung in allen Bereichen die Fehlerzahl verringern läßt, während dies bei den anderen beiden Klassifikatoren für $M > 25$ kaum mehr möglich ist. So ist die Fehlerrate für $M = 40$ Merkmale bei dem Quadratmittelklassifikator nur etwa halb so groß, wie die des Normalverteilungsklassifikators und etwa ein Drittel so groß wie die des abstandsmessenden Klassifikators.

Als weiteres Maß für die Systemleistung verwenden wir die Rückweisungsrate, die notwendig ist, um eine Fehlerrate von 10^{-3} zu erhalten. In Bild 87 ist $g_{\epsilon=10^{-3}}$ für die drei Systeme über M abgetragen. Aus dem Diagramm ist erkennbar, daß für alle Werte $M > 20$ der Quadratmittelklassifikator weniger Zeichen zurückweisen muß, um diese Fehlerrate zu erreichen. Wiederum fällt der flache Verlauf der Kurve des Normalverteilungsklassifikators für $M > 30$ auf. Auch hier zeigt sich, daß die Modellierung des Zufallsprozesses $\{w/k\}$ durch die Normalverteilung für derart große Dimensionen M des Merkmalsraumes nicht mehr ausreichend gut gelingt. Dies spiegelt sich in dem Sättigungseffekt dieser Kurve wieder. Dagegen läßt sich offenbar die Schätzung der Rückschlußwahrscheinlichkeit $p(k/w)$ durch den Quadratmittelklassifikator durch Vergrößern der Koeffizientenzahl stets verbessern.

In Bild 88 ist der Adaptionaufwand der Systeme, der mit der Rechnerkernzeit gemessen wird, die zur Bestimmung der Koeffizienten der Schätzpolynome notwendig war, abgetragen über der Zahl M der verwendeten Merkmale. Das Bild zeigt deutlich, wie sich die verschiedenen Systeme hinsichtlich des Adaptionaufwandes unterscheiden: dieser ist für den Quadratmittelklassifikator, auch bei einer

6 BIT Quantisierung der Merkmale w_i doch erheblich größer als bei den beiden anderen Systemen. Damit bietet sich der Einsatz von Normalverteilungsklassifikatoren, - die, wie aus den Experimenten hervorging, eine bessere Leistung erbrachten als der abstandsmessende Klassifikator - dann an, wenn der Anteil der Adaptionskosten an den Gesamtkosten des Erkennungssystems sehr hoch sind, gleichzeitig aber keine zu großen Forderungen an die Systemleistung gestellt werden.

Neben dem Vergleich der Leistung der verschiedenen Systeme ist nun interessant, die Koeffizienten der jeweiligen Schätzpolynome dieser Klassifikatoren zu betrachten, um daraus Schlüsse ziehen zu können, in welcher Art und Weise die Punkte im Merkmalsraum den verschiedenen Klassen zugeordnet werden. Wir nehmen zunächst auf die für alle Systeme einheitliche mathematische Form der Schätzfunktion (Gleichung (97)) Bezug.

Für den Normalverteilungsklassifikator und abstandsmessenden Klassifikator ist das Zentrum der quadratischen Form der Klassenschwerpunkt. Die Kernmatrix ist die Inverse der klassenspezifischen Kovarianzmatrix und damit positiv semidefinit. Alle Eigenwerte sind somit größer oder gleich Null. Der Wert d_k erreicht im Klassenschwerpunkt sein Maximum, nämlich $d_k = \mu_k$, und nimmt daher in allen Richtungen des Merkmalsraumes ab. Damit ist unmittelbar die obere Grenze von d_k gegeben. Die geometrischen Orte von Konstanten d_k sind Hyperellipsoide.

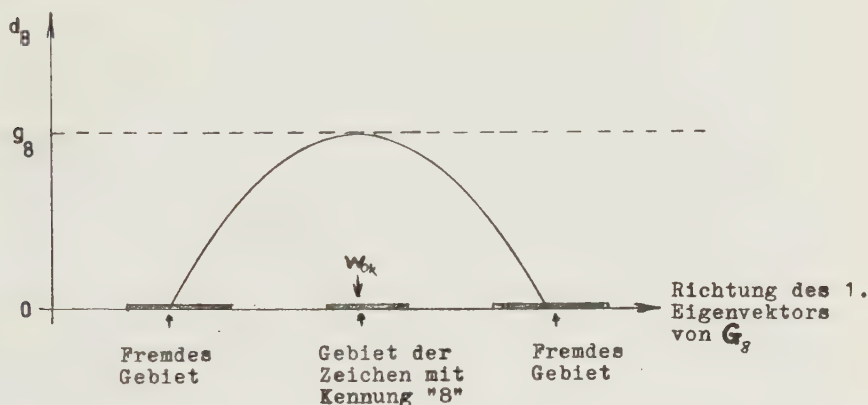
Derartige grundsätzlichen Aussagen sind nun für den Quadratmittelklassifikator nicht möglich. Möchte man bei diesem Klassifikator solche Verständnisfragen beantworten, so müssen die Koeffizienten eines speziellen Klassifikators vorliegen. Die abgeleiteten Schlüsse gelten dann nur für dieses betrachtete System.

Die Berechnung der Kernmatrizen G_k , der Vektoren w_{0k} und der Konstanten g_k , sowie der Eigenvektoren und Eigenwerte von G_k wurde durchgeführt für den Quadratmittelklassifikator mit dem in $M = 40$ Merkmalen vollständig quadratischen Ansatzpolynom (HO4HTQ). Dabei zeigte sich, daß keine der 10 Kernmatrizen definit ist, d. h., bei jeder Matrix G_k existieren positive und negative Eigenwerte gleichzeitig. Damit ist über d_k eine Abschätzung wie beim Normalverteilungsklassifikator nicht möglich. Entfernt man sich vom Zentrum w_{0k} der quadratischen Form, so kann sich der Wert d_k vergrößern oder verkleinern. Erstaunlich ist, daß die betragsgrößten Eigenwerte von G_k stets negativ sind. Eine Ausnahme hierbei ist nur die Kernmatrix der Kennung "8", deren betragsgrößter Eigenwert positiv ist (Bild 89). Von den insgesamt 40 Eigenwerten waren bei jeder Klasse in etwa die Hälfte positiv (Bild 90). Da keiner der Eigenwerte den Wert 0 besitzt, existieren Ebenen im Merkmalsraum, in denen die Schnitte der Flächen $d_k = \text{const}$ Ellipsen und Ebenen, in denen diese Schnittkurven Hyperbeln sind.

In Bild 91 sind die sich ergebenden Konstanten g_k für alle Klassen aufgelistet. Bemerkenswert an der Werteverteilung ist, daß nahezu sämtliche g_k sehr klein sind. Ein Merkmalsvektor $w = w_{0k}$ wird damit als nicht zu der Klasse k zugehörig geschätzt. Damit liegt w_{0k} weit außerhalb des Gebietes der Klasse k . Eine Ausnahme bildet dabei nur die Kennung "8", deren Wert $g_8 = 0.58$ relativ groß ist.

Aus früheren Experimenten wissen wir, daß die Klasse "8" insofern ausgezeichnet ist, als ihr Klassengebiet etwa in der Mitte des überhaupt von der Stichprobe beanspruchten Gebietes im Merkmalsraum liegt. Die Zeichen der Klasse "8" sind dem "mittleren Zeichen" der Stichprobe (Schwerpunktvektor $\bar{v}$) am ähnlichsten. Dies ermöglicht dann die folgende Interpretation der vorliegenden Situation im Merkmalsraum.

Hierzu betrachten wir - für jede Klasse getrennt-- die Richtung des Eigenvektors von $\mathbf{G}_k$, die zu dem betragsgrößten Eigenwert von $\mathbf{G}_k$ gehört. Für die Zeichen der Kennung "8" liegt dann in dieser Merkmalsrichtung die skizzierte Situation vor:



Da der zugehörige Eigenwert positiv ist, muß die Parabel nach unten geöffnet sein. Das Gebiet der "8" ist von den Gebieten der anderen Klassen umgeben. Diese fremden Gebiete kommen in dem Bereich zu liegen, in dem die Parabel einen Nulldurchgang besitzt, d. h., die Schätzung der Klassenzugehörigkeit liegt um den Wert $d_8 = 0$. Der Punkt $\mathbf{w} = \mathbf{w}_{0k}$ liegt im Klassengebiet. An dieser Stelle besitzt die Parabel ihren Scheitelpunkt; die geschätzte Klassenzugehörigkeit ist hier $d_8 = g_8 = 0.58$.

Für die Zeichen aller anderer Klassen liegt eine andere Situation vor.



Da der zugehörige Eigenwert negativ ist, muß die Parabel nach oben geöffnet sein. Das Gebiet der eigenen Klasse k liegt etwas abseits von dem Gebiet aller fremden Klassen. Im eigenen Klassengebiet nimmt d_k Werte um 1 an, während d_k im fremden Klassengebiet Werte um 0 annimmt. Für den Punkt $\mathbf{w} = \mathbf{w}_{0k}$ ergibt sich dabei entlang dieser Gerade der minimale Wert, nämlich $d_k = g_k$ (Bild 91).

6. Zusammenfassung

In der Arbeit werden verschiedene Erkennungssysteme betrachtet, die aus zwei, nach unterschiedlichen Kriterien getrennt optimierten Stufen bestehen, der Merkmalsextraktionsstufe und der Klassifikationsstufe. Die Aufgabenstellung ist gegeben durch den zeichenerzeugenden Zufallsprozess $\{\mathbf{v}\}$, der durch eine umfangreiche Stichprobe repräsentiert wird. Elemente der Stichprobe sind dabei die Merkmalsvektoren $\mathbf{v}$, durch die die binär quantisierten Rasterbilder der zu klassifizierenden Zeichen beschrieben werden. Um über diese Rasterbilder der Zeichen verfügen zu können, wurden die Zeichen abgetastet und dann einem Zeichenvorbereitungsprozess zugeführt, der sie auf eine normierte Form brachte.

Die Zeichenvorbereitungsprozeduren, deren Effektivität die Leistung des Erkennungssystems auch entscheidend bestimmen, werden hier nicht betrachtet. Es wird die Verfügbarkeit der Stichprobe quantisierter und normierter Rasterbilder $\{\mathbf{v}\}$ vorausgesetzt.

Die erste Stufe des Erkennungssystems hat die Aufgabe, die Zeichen möglichst gut zu beschreiben. Um dies zu erreichen, wird der N-dimensionale Zufallsprozess $\{\mathbf{v}\}$ in einem neuen Koordinatensystem mit wesentlich reduzierter Zahl M von Dimensionen beschrieben. Diese Beschreibung gelingt im allgemeinen nun nicht mehr fehlerfrei. Dies bedeutet, daß jedem Zeichen ein Fehlervektor zugeordnet werden kann, dessen über alle Zeichen gemitteltetes Betragsquadrat möglichst klein werden soll. Diese Forderung nach Minimierung des mittleren Betragsquadrats des Fehlervektors ist das Optimierungskriterium der ersten Stufe. Mittelpunkt und Richtung der Koordinatenachsen des neuen Systems werden so festgelegt, daß dieses Optimierungskriterium erfüllt wird. Die neuen Merkmale, mit denen die Zeichen nun beschrieben werden, sind die Koordinaten der Punkte im neuen Koordinatensystem, durch die die Zeichen im Merkmalsraum repräsentiert werden. Diese Abbildung des N-dimensionalen Merkmalsvektors $\mathbf{v}$ in den M-dimensionalen Merkmalsvektor $\mathbf{w}$ wird auch als Hauptachsen-transformation bezeichnet. Wichtig bei der Bestimmung des neuen Koordinatensystems ist, daß ausschließlich der Zufallsprozess $\{\mathbf{v}\}$ und nicht die Prozesse $\{\mathbf{v}/k\}$ betrachtet werden. Eine Kenntnis der Zugehörigkeit der Zeichen zu einer Klasse ist nicht notwendig. Damit wird eine vollständige Entkoppelung der beiden Stufen erreicht.

Es wird mit mehreren anschaulichen Methoden dargelegt, inwieweit die Beschreibung des Zufallsprozesses $\{\mathbf{w}/k\}$ durch einfache und hantierbare Modelle, insbesondere der Normalverteilung sinnvoll und möglich ist. Hierzu werden eindimensionale Randverteilungen ausgemessen. Die Projek-

tion der Punktmenge einer Klasse auf ausgesuchte Ebenen geben einen flächenhaften Eindruck der Verteilung wieder. Eine weitere Möglichkeit, Modellprozess mit dem gegebenen Prozess zu vergleichen, ist die Betrachtung der Verteilung des verallgemeinerten Abstandsmaßes $Q_k = (\mathbf{w} - \bar{\mathbf{w}}_k)^T \mathbf{K}_k^{-1} (\mathbf{w} - \bar{\mathbf{w}}_k)$. Unter Annahme von normalverteiltem $\{\mathbf{w}/k\}$ ergibt sich dabei eine χ^2 Verteilung. Die Gegenüberstellung von dieser Verteilung mit der gemessenen Verteilung von Q_k vermittelt einen Eindruck, inwieweit die M-dimensionalen Verteilungen übereinstimmen. Durch Integration kann schließlich auch der Anteil aller Zeichen außerhalb eines gegebenen Ellipsoides bestimmt werden, und mit den gemessenen Werten verglichen werden.

Optimierungskriterium der zweiten Stufe ist die Minimierung des Risikos. Die Einführung dieses Kriteriums ergibt sich unmittelbar aus der statistischen Betrachtungsweise, daß alle Entscheidungen des Systems mit Kosten bewertet werden und die mittleren Kosten möglichst klein gehalten werden sollen. Damit ist die Struktur der zweiten Stufe gegeben: es gilt die Rückschlußwahrscheinlichkeit $p(k/\mathbf{w})$ zu schätzen. Nun gibt es verschiedene Möglichkeiten, aus dem Merkmalsvektor $\mathbf{w}$ diese Rückschlußwahrscheinlichkeit zu schätzen. Drei werden davon betrachtet: bei der ersten Methode wird angenommen, es liege ein normalverteilter Prozess $\{\mathbf{w}/k\}$ vor. Die Parameter der Verteilung sind aus der Stichprobe zu bestimmen. Damit läßt sich $p(k/\mathbf{w})$ unmittelbar berechnen (Normalverteilungsklassifikator). Bei der zweiten Methode wird zusätzlich vereinfachend angenommen, die Determinanten der klassenspezifischen Kovarianzmatrizen sind gleich. Damit ist ebenfalls eine Schätzung der Rückschlußwahrscheinlichkeit möglich (abstandsmessender Klassifikator). Bei der dritten Methode wird $p(k/\mathbf{w})$ mit einem in den Merkmalen w_i quadratischen Polynom geschätzt, bei dem die Koeffizienten so berechnet werden, daß der mittlere quadratische Fehler zwischen Soll- und Zielvek-

tor minimal wird (Quadratmittelklassifikator).

Diesen Klassifikatoren gemeinsam ist das in den Merkmalen vollständig quadratische Schätzpolynom, das mathematisch einheitlich als quadratische Form geschrieben werden kann. Bestimmende Parameter sind dabei Kernmatrix und Zentrum. Mit dieser Darstellung der Schätzfunktionen ist es für alle drei betrachteten Systeme möglich, Aussagen über den Verlauf der geometrischen Orte konstanter Schätzung im Merkmalsraum zu machen. Grundsätzlich sind diese geometrischen Orte für den Normalverteilungsklassifikator und abstandsmessenden Klassifikator M-dimensionale Hyperellipsoide, während beim Quadratmittelklassifikator eine derartige Beschränkung nicht gegeben ist: als Schnittfiguren können hier sämtliche Kegelschnitte entstehen. Um eine anschauliche Vorstellung dieser Kurven zu vermitteln, wurden Klassifikatoren mit nur zwei Merkmalen berechnet. Die Situation im Merkmalsraum ist dann auf der Zeichenebene darstellbar. Die Schlüsse, die aus dem Verlauf der geometrischen Orte konstanter Schätzung gezogen werden, tragen zum Verständnis des Klassifizierungsprinzips bei.

Entscheidend für die Beurteilung eines Erkennungssystems ist letztlich seine praktische Anwendbarkeit. Dabei interessiert insbesondere, welche Leistung ein System bei welchem Aufwand erreicht. Ebenfalls wichtig ist die Ermittlung derjenigen Leistung, die auch bei Erhöhung des Aufwandes nicht mehr verbessert werden kann.

Diese Leistungsmessungen wurden unter Verwendung der umfangreichen Handblockschrift-Stichprobe durchgeführt. Für alle drei Klassifikatoren wurden dabei Fehlerraten und Rückweisungs-raten in Abhängigkeit von Zurückweisungsschwellen bestimmt. Als Maß für den Klassifikationsaufwand wurde die Zahl der pro Klasse abzuspeichernden Koeffizienten eingeführt, die benötigt werden, um die Klassenzugehörigkeit

zu schätzen. Adaptionaufwand wird mit der zur Berechnung des Klassifikators notwendigen Rechnerkernzeit gemessen.

Bei den Experimenten zeigt sich die deutliche Überlegenheit des Normalverteilungsklassifikators gegenüber dem abstandsmessenden Klassifikator. Es war zwar bekannt, daß sich die Determinanten der klassenspezifischen Kovarianzmatrizen unterscheiden, und damit auch bekannt, daß die Hypothese, aufgrund der der abstandsmessende Klassifikator konzipiert ist, nur in grober Näherung erfüllt ist. Derartig starke Auswirkungen auf das Klassifizierungsergebnis konnten aber nicht vorhergesehen werden. Der Normalverteilungsklassifikator ist bei Verwendung von wenigen Merkmalen leistungsfähiger als der Quadratmittelklassifikator. Allerdings stagniert die Leistung des Normalverteilungsklassifikators ab 30 Merkmalen, während die Leistung des Quadratmittelklassifikators bei einer Erhöhung der Zahl der verwendeten Merkmale von 30 auf 40 noch erheblich gesteigert werden konnte. Damit zeigte sich, daß der Quadratmittelklassifikator, der 40 Merkmale auswertet, deutlich leistungsfähiger als der Normalverteilungsklassifikator ist.

Das Stagnieren der Leistung des Normalverteilungsklassifikators bei $M = 30$ und mehr verwendeten Merkmalen deutet darauf hin, daß der Zufallsprozess $\{W/k\}$ mit derartig großer Zahl M von Dimensionen nicht mehr genügend genau durch eine Normalverteilung modellierbar ist.

Zeichnet sich der Quadratmittelklassifikator einerseits durch die hohe Leistungsfähigkeit aus, so muß andererseits auch der große Adaptionaufwand erwähnt werden, der bei allen Verfahren gegeben ist, die sämtliche statistische Information über den zeichenerzeugenden Zufallsprozess erst während der Adaptionphase gewinnen. Für diesen Klassifikator wurden daher Experimente durchgeführt, die zeigten, daß durch die Quantisierung der Merkmale eine Beschleunigung des Adaptionprozesses

möglich ist, wobei kaum eine Leistungseinbuße in Kauf genommen werden mußte. Schließlich konnte die Fehler- und Rückweisungsrate des Quadratmittelklassifikators durch einen Iterationsprozess verbessert werden. Hierbei werden schlecht erkannte Zeichen in einem neuen Adaptionsprozess verstärkt berücksichtigt. Damit ist ein fehlerfreies Erkennen der gesamten Handblockschrift-Stichprobe bei Zurückweisung von nur 0.5% aller Zeichen möglich.

Die Ergebnisse dieser Arbeit fließen unmittelbar in die Entwicklung von Lesemaschinen ein. Das hier ausführlich diskutierte Konzept des zweistufigen Erkennungssystems, bestehend aus Hauptachsentransformation und Quadratmittelklassifikator zweiten Grades, ist Grundlage dafür, die Leistung eines direkten Quadratmittel-Polynomklassifikators zweiten Grades, wie er in schnellen Beleglesemaschinen eingesetzt wird, wesentlich zu steigern /02/. Bei den direkten Quadratmittelklassifikatoren sind die Argumente der Schätzfunktion direkt die Merkmale v_i bzw. die Merkmalsverknüpfungen $v_i v_j$. Mit den in /02/ beschriebenen Verfahren konnte die Lesesicherheit dieser Maschinen, durch Bestimmung einer Auswahl von besonders effektiven Polynomtermen $v_i v_j$, erheblich erhöht werden. Änderungen an der Hardware waren hierzu nicht erforderlich.

Dieses zweistufige System (HAT und Quadratmittelklassifikator), dessen Leistungsfähigkeit am Beispiel des Handblockschrift-Erkennungsproblem hier gemessen wurde, wird ebenfalls erfolgreich zur automatischen Erkennung von gesprochenen Befehlsworten eingesetzt.

Literaturverzeichnis

- /A1/ H. C. Andrews
Multidimensional Rotations in Feature Selection
IEEE Transactions on Computers, Vol. C-20, (1971)
No. 9, pp. 1045-1051
- /A2/ H. C. Andrews
Some Unitary Transformations in Pattern
Recognition and Image Processing
IFIP-Congress 1971, Ljubljana
Foundations of Information Processing
- /A3/ A. Abramowitz, I. A. Stegun
Handbook of Mathematical Functions
Nat. Bureau of Standards, 1965
- /A4/ H. C. Andrews
Introduction to Mathematical Techniques in
Pattern Recognition
Wiley-Interscience, 1972
- /B1/ J. F. Böhme
Über Umformung von Mustern zur Klassifizierung
In: Arbeitsberichte des Instituts für mathema-
tische Maschinen und Datenverarbeitung
Friedrich Alexander Universität Erlangen
Nürnberg, Band 2, Nummer 5, August 1969
- /B2/ R. S. Burington
Handbook of Mathematical Tables and Formulas
Mc Graw Hill, New York 1965
- /B3/ D. Becker, J. Schürmann
Zur verstärkten Berücksichtigung schlecht erkenn-
barer Zeichen der Lernstichprobe
Wiss. Berichte AEG Telefunken 45 (1972) 1/2, S. 97-105

- /B4/ H. Bleichrodt, W. Schwerdtmann
Sequentielle und nichtsequentielle Klassifikation
bei einem Erkennungssystem für handgeschriebene
Zeichen
NTZ 28 (1975), H. 10, S. 354 - 355
- /C1/ Y. T. Chien, K. S. Fu
On the generalized Karhunen-Loeve expansion
IEEE Transactions on Inform. Theory,
vol. IT-13 (1967), No. 3, pp. 518 - 520
- /C2/ T. W. Calvert
Nonorthonormal Projections for Feature
Extraction in Pattern Recognition
IEEE Transactions on Computers, vol. C-19 (1970),
No. 5, pp. 447 - 452
- /C3/ C. H. Chen
Theoretical Comparison of a Class of Feature
Selection Criteria in Pattern Recognition
IEEE Transactions on Computers, vol. C-20 (1971),
No. 9, pp. 1054 - 1056
- /D1/ R. O. Duda, P. E. Hart
Pattern Recognition Techniques, Butterworth
London 1973
- /F1/ F. Fukunaga, W. L. Koontz
Application of the Karhunen-Loeve Expansion to
Feature Selection and Ordering
IEEE Transactions on Computers, vol. C-19 (1970),
No. 4, pp. 311 - 318
- /F2/ K. S. FU, PYUNG JUNE MIN, T. J. LI
Feature Selection in Pattern Recognition
IEEE Transactions on Systems Science and
Cybernetics, vol. SSC-6 (1970), No. 1, pp. 33 - 39

- /F3/ D. H. Foley, J. W. Sammon
An Optimal Set of Discriminant Vectors
IEEE Transactions on Computers,
vol. C-24 (1975), No. 3, pp. 281 - 289
- /F4/ Firmeninterne Berichte
- /F5/ K. Fukunaga
Introduction to Statistical Pattern Recognition
Academic Press, London, New York, 1972
- /G1/ H. Giebel
Optimierung mustererkennender Schichtstrukturen
durch Merkmalssynthese
Dissertation, TU München 1973
- /G2/ A. Güdesen
Vorverarbeitung handgeschriebener gerasterter
Ziffern
Fachtagung "Cognitive Verfahren und Systeme"
Hamburg 11. - 13. April 1973
Springer Verlag, Berlin, Heidelberg, New York,
S. 209 - 222
- /H1/ D. J. Hunt
A Feature Extraction Method for the Recognition
of Handprinted Characters
Proceedings of the Conference: Machine Perception
of Patterns and Pictures, pp. 28 - 33
Teddington, Middlesex 12. - 14. April 1972
- /K1/ P. Krause, W. Schwerdtmann, D. Paul
Two Modifications of a Recognition System with
Pattern Series Expansion and Bayes Classifier
Proceedings of the II Int. Joint Conference on
Pattern Recognition, Aug. 13 - 15, 1974, Copenhagen
pp. 215 - 219

- /K2/ L. Kanal
Patterns in Pattern Recognition: 1968 - 1974
IEEE Transactions on Inform. Theory,
vol. IT-20 (1974), No. 6, pp. 697 - 722
- /K3/ J. Kittler, P. C. Young
A New Approach to Feature Selection Based on the
Karhunen-Loeve Expansion
Pattern Recognition, Pergamon Press 1973
vol. 5 pp. 335 - 352
- /K4/ B. R. Kowalsky, G. F. Bender
An Orthogonal Feature Selection Method
Pattern Recognition, Pergamon Press 1976
vol. 8, pp. 1 - 4
- /K5/ E. Kreyszig
Statistische Methoden und ihre Anwendungen
Vandenhoeck u. Ruprecht, Göttingen, 1967
- /K6/ V. A. Kovalevsky
Character Readers and Pattern Recognition
Spartan 1968, New York
- /K7/ L. Kanal, B. Chandrasekaran
On Dimensionality and Sample Size in Statistical
Pattern Classification
Pattern Recognition, Pergamon Press 1971
vol. 3, pp. 225 - 234
- /L1/ C. L. Lawson, R. J. Hanson
Solving Least Square Problems
Prentice Hall, Inc, Englewood Cliffs, New Jersey,
1974

- /L2/ P. A. Lachenbruch, M. R. Miekey
Estimation of Error in Discriminant Analysis
Technometrics (1968), vol. 10, pp. 1 - 11
- /M1/ G. Meyer-Brötz, J. Schürmann
Probleme der Mustererkennung bei der automatischen Briefverteilung
Wiss. Berichte AEG-Telefunken 48 (1075) 2/3,
S. 85 - 90
- /M2/ P. H. Müller
Wahrscheinlichkeitsrechnung und mathematische Statistik
Lexikon, Akademie-Verlag, Berlin 1970
- /M3/ W. S. Meisel
Computer-oriented Approaches to Pattern Recognition
Academic Press, New York 1972
- /M4/ G. Meyer-Brötz, J. Schürmann
Methoden der automatischen Zeichenerkennung
Oldenbourg, München Wien 1970
- /N1/ H. Niemann
Begründung und Anwendung einer Theorie zur quantitativen Beschreibung und Erkennung von Mustern
Dissertation, TU Hannover 1969
- /N2/ H. Niemann
Mustererkennung mit orthonormaler Reihenentwicklung
NTZ 23 (1970), H. 6, S. 308 - 313

- /N3/ H. Niemann
Methoden der Mustererkennung
Akademische Verlagsgesellschaft Frankfurt am Main
1974
- /N4/ H. Niemann
Die Erzeugung von Mustern mit vorgegebenen statistischen Eigenschaften
NTZ 24 (1971), H. 5, S. 273 - 275
- /N5/ N. J. Nilsson
Learning machines
Mc Graw Hill, New York 1965
- /O1/ R. Ott, J. Schürmann
Zum Problem der Merkmalsreduzierung bei der Zeichenerkennung
Wiss. Berichte AEG-Telefunken 47 (1974) 3/4,
S. 100 - 110
- /O2/ R. Ott
Construction of Quadratic Polynomial Classifiers
Proc. III I. J. C. on Pattern Recognition,
Nov. 8 - 11, 1976, Coronado, California,
pp. 161 - 165
- /R1/ N. Ahmed, K. R. Rao
Orthogonal Transforms for Digital Signal Processing
Springer Verlag Berlin, Heidelberg, New York 1975
- /R2/ C. R. Rao
Linear Statistical Inference and its Applications
John Wiley & Sons Inc, New York, London, Sydney
1973 (2. Auflage)

- /S1/ J. Schürmann
Über die Anwendung der Regressionsanalyse auf das
Zeichenerkennungsproblem
Dissertation, TU Berlin, 1968
- /S2/ J. Schürmann
Experimenteller Vergleich verschiedener Schwer-
punkt-Zentrierverfahren für die Zeichenerkennung
NTZ 21 (1968), H. 6, S. 312 - 315
- /S3/ J. Schürmann
Bildvorbereitung für die automatische Zeichen-
erkennung
Wiss. Berichte AEG-Telefunken 47 (1974) 3/4,
S. 90 - 99
- /S4/ J. Schürmann, L. Vogt
Vergleich linearer und quadratischer Klassifika-
toren am Beispiel der Schriftart OCR-A
NTZ 26 (1973), H. 6, S. 269 - 274
- /S5/ J. Schürmann
Über systematisch konstruierte nichtlineare
Klassifikatoren für die Handblockschrift-Erkennung
Elektronische Rechenanlagen 13 1971, H. 6,
S. 250 - 260
- /S6/ J. Schürmann
Zur Zurückweisung zweifelhafter Zeichen
NTZ 26 (1973), H. 3, S. 137 - 144
- /S7/ J. Schürmann
Zeichenerkennung durch Rechnersimulation
in: Zeichenerkennung durch biologische und
technische Systeme
Springer Verlag, Berlin Heidelberg New York 1971
S. 176 - 185

- /S8/ J. Schürmann, R. Ott, P. Hauff
Zur Konstruktion quadratischer Klassifikatoren
mit direktem Polynomansatz
Wiss. Berichte AEG-Telefunken 49 (1976) 1/2,
S. 42 - 50
- /S9/ J. Schürmann
Multifont Word Recognition System with Application
to Postal Address Reading
Proc. III I. J. C. on Pattern Recognition,
Nov. 8 - 11, 1976, Coronado, California,
pp. 658 - 662
- /T1/ S. Tornita, S. Noguchi, J. Oizumi
On Evaluation of Handwritten Characters by
Karhunen-Loeve Orthogonal System
Proc. 3rd Hawaii Int. Conf. on System Sci.
Part I, 1970, pp. 501 - 504
- /T2/ C. W. Therrien
Eigenvalue Properties of Projection Operators
and their Application to the Subspace Method
of Feature Extraction
Prepared for Electronic System Division
Ballistic Missile Defense Programm Office
14. August 1974
- /T3/ C. W. Therrien
A Generalized Approach to Linear Methods of Feature
Extraction
In: Proceedings of the Conference on Computer
Graphics, Pattern Recognition and Data Structure
May 14 - 16, 1975, pp. 312 - 315

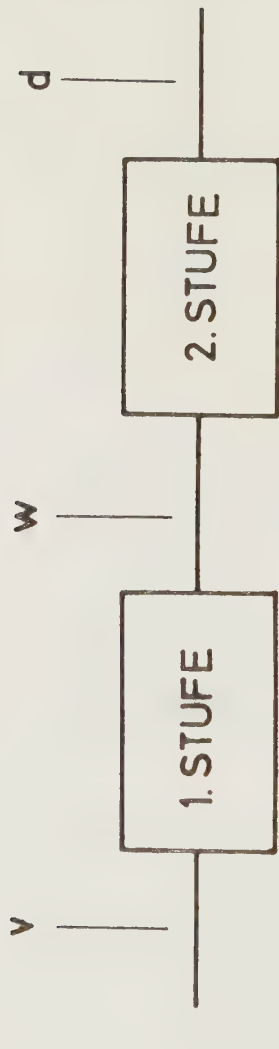
- /T4/ J. T. Tou
Feature Selection for Pattern Recognition Systems
In: S. Watanabe ed. 'Methodologies of Pattern
Recognition'
Academic Press, New York, 1969
- /U1/ J. R. Ullmann
A Review of Optical Pattern Recognition Techniques
Opto electronics 6 (1974) pp. 319 - 332
- /W1/ S. Watanabe
Karhunen-Loeve Expansion and Factor Analysis
Transaction of the Fourth Prague Conference on
Information Theory, pp. 635 - 660
Prag 1967
- /W2/ S. Watanabe
Feature Compression
J. T. Tou ed 'Advances in Information System
Sciences' Vol. 3, Plenum Press. New York 1970
- /W3/ A. W. Witney
A Direct Method of Nonparametric Measurement
Selection
IEEE Transactions on Computers, vol. C-20 (1971),
No. 9, pp. 1100 - 1103
- /W4/ S. Watanabe
Frontiers of Pattern Recognition
Academic Press 1972
- /Y1/ S. S. Yau, J. M. Garnett
Least-Mean-Square Approach to Pattern Recognition
in: Frontiers of Pattern Recognition
Academic Press, New York, 1972

/Z1/

R. Zurmühl

Matrizen und ihre technische Anwendungen

Springer Verlag, Berlin/Göttingen/Heidelberg 1964



Merkmals-
vektor $\mathbf{v}$

$\text{Dim } \{\mathbf{v}\} = N \approx 200$

Rasterbild nach Bild-
vorbereitung

Merkmals-
vektor $\mathbf{w}$

$\text{Dim } \{\mathbf{w}\} = M \approx 40$

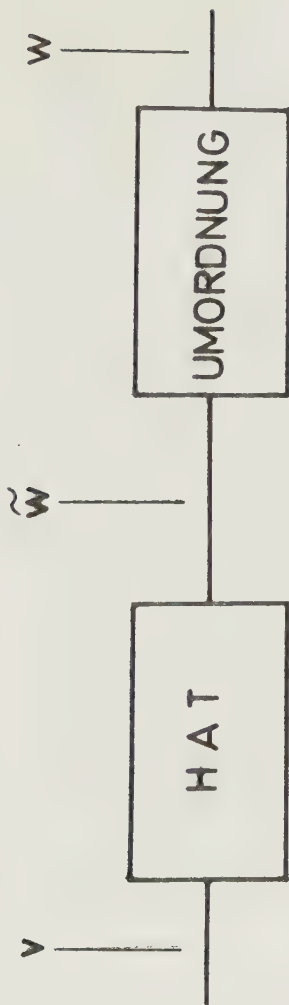
w_i sind die
neuen Merkmale

Schätzvektor $\mathbf{d}$

$\text{Dim } \{\mathbf{d}\} = K$

d_k sind die Ent-
scheidungsmaße

Bild 1. Das zweistufige Zeichenerkennungssystem



$$\text{Dim } \{v\} = N$$

$$\text{Dim } \{\tilde{w}\} = N$$

$$\text{Dim } \{w\} = N$$

w_i geordnet nach
Kriterium Rekon-
struktion ΔR_1^2

w_i geordnet nach
Kriterium Klassi-
fikation ΔS_1^2

Bild 2. Umordnen der Merkmale w_i hinsichtlich des Klassifizierungskriteriums

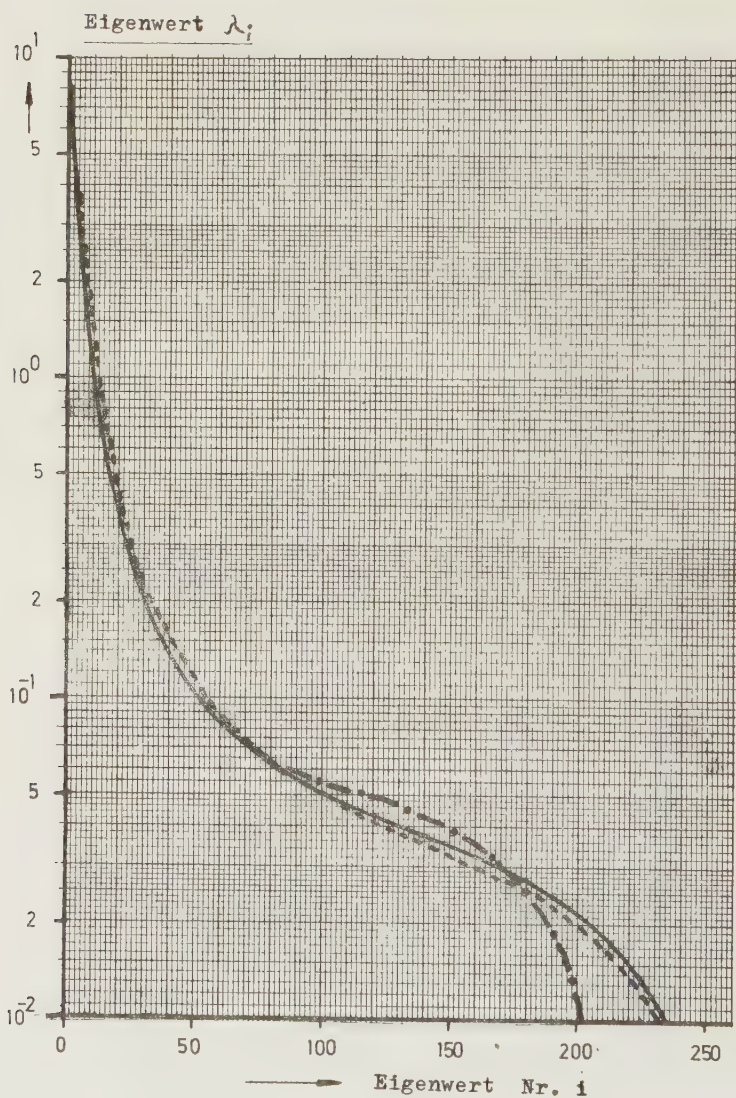


Bild 3. Verteilung der Eigenwerte der Kovarianzmatrizen verschiedener Stichproben.

— A2 - - - A1 -•-•- H0

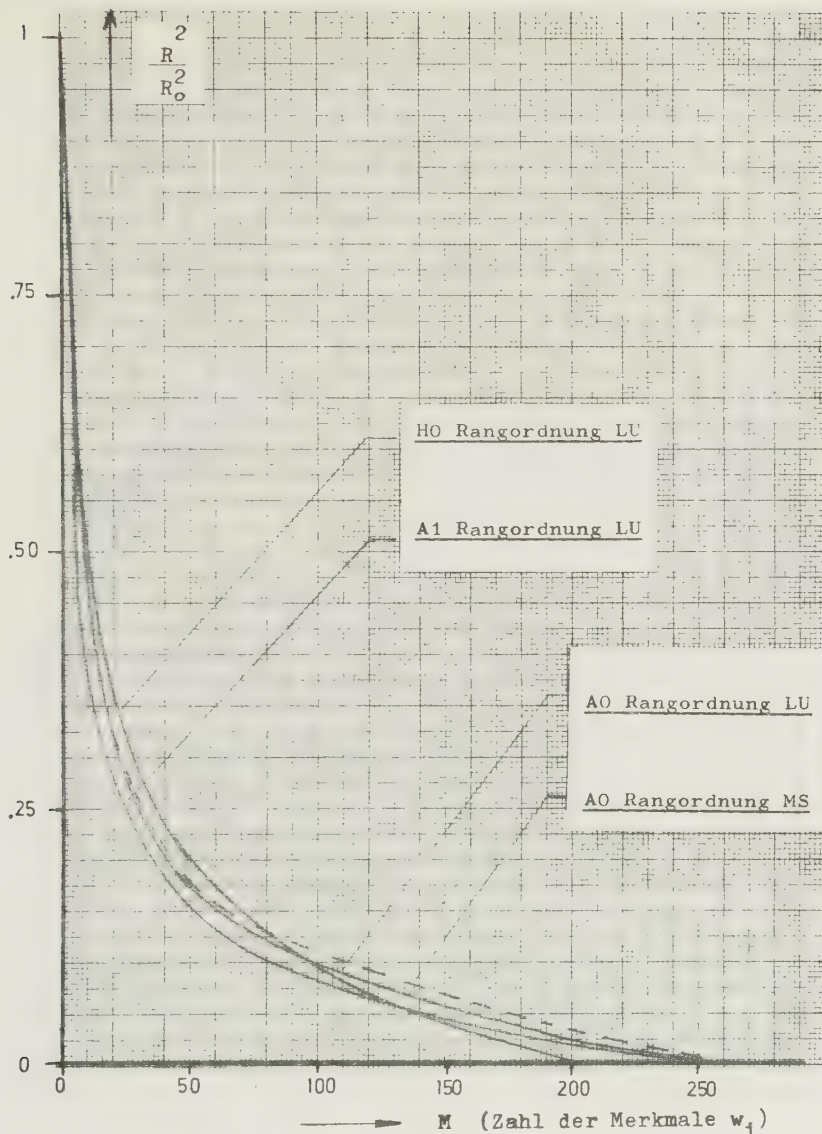


Bild 4. Normierter Rekonstruktionsfehler R^2/R_0^2 , abgetragen über der Zahl der Merkmale w_1 . Die Kurve R^2/R_0^2 für die Stichprobe A2 liegt zwischen den Kurven der Stichprobe AO und A1.

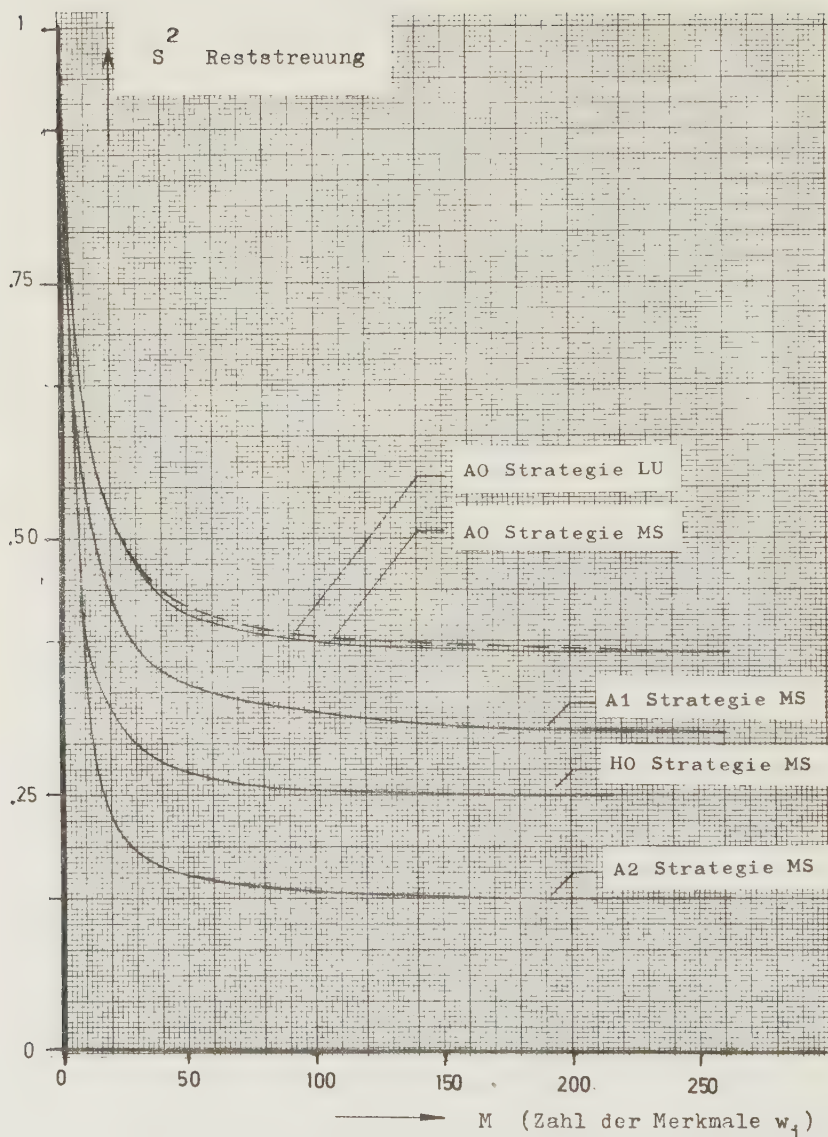


Bild 5. Die Reststreuung S^2 des linearen Quadratmittelklassifikators abgetragen über der Zahl M der in die Schätzgleichung aufgenommenen Variablen w_i

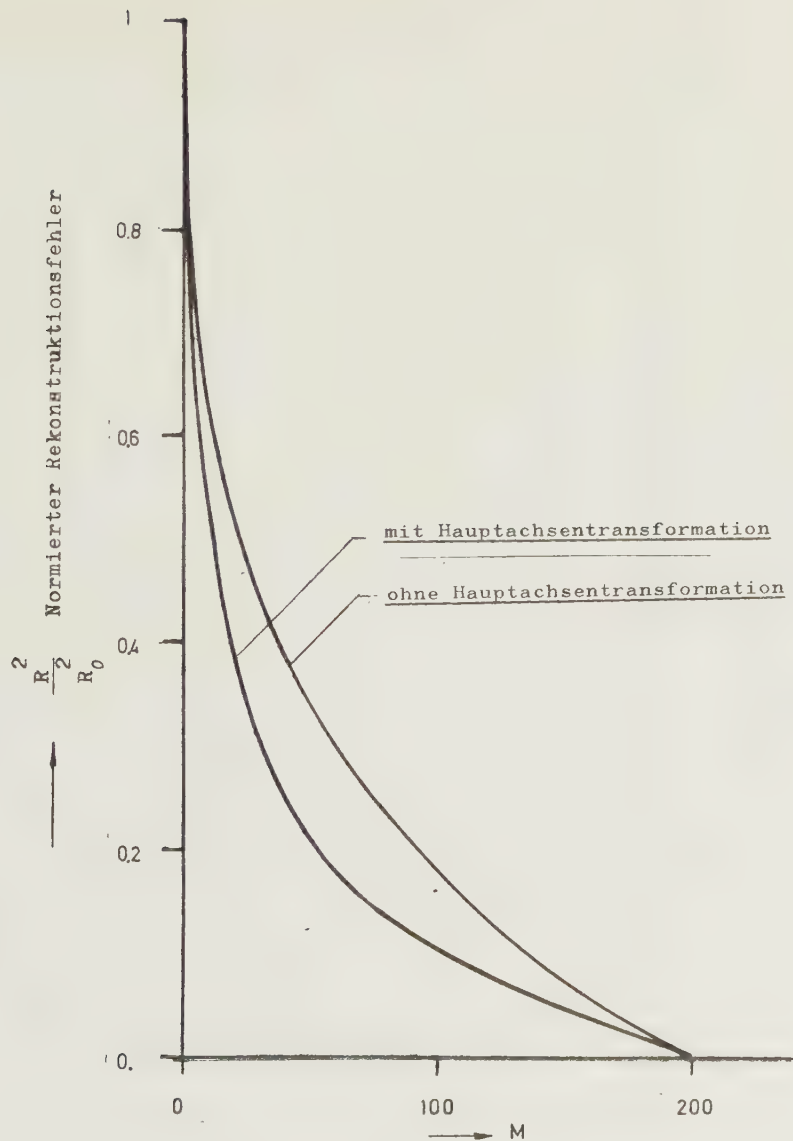


Bild 6. Der normierte Rekonstruktionsfehler R^2/R_0^2 in Abhängigkeit von der Zahl M weiterverwendeter Dimensionen mit und ohne Hauptachsentransformation. Die Merkmalsauswahl erfolgte nach der MS-Rangordnung.

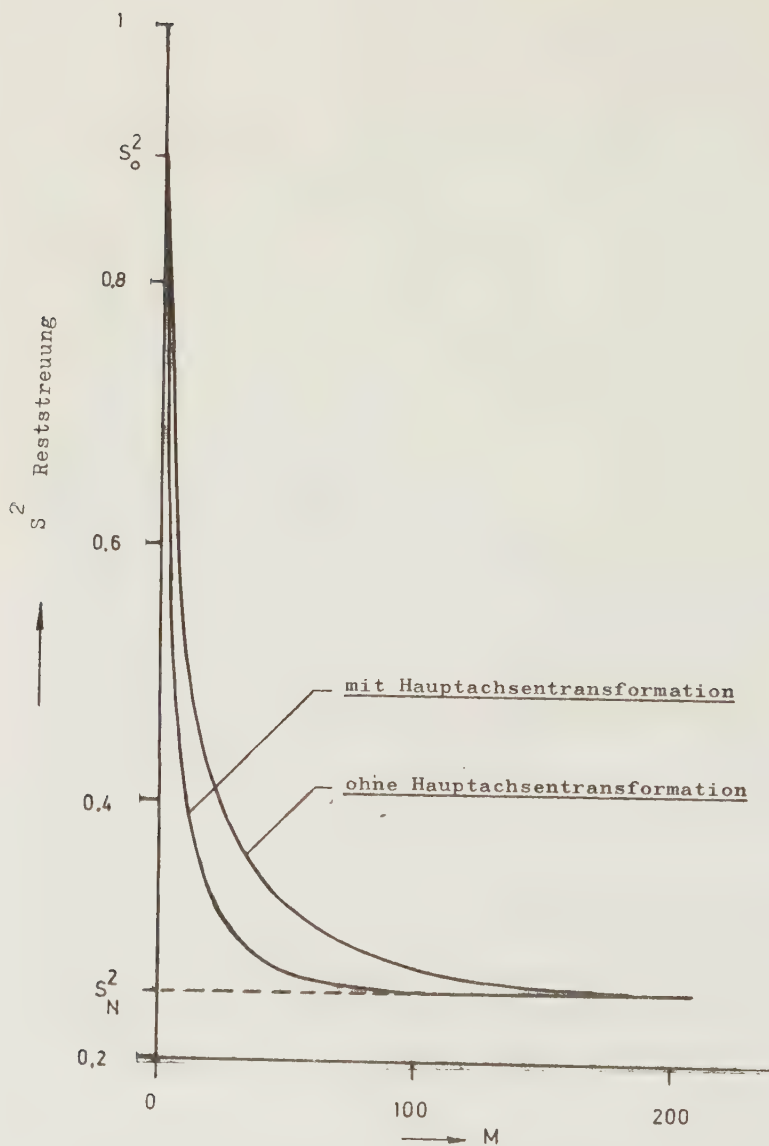


Bild 7. Die Reststreuung S^2 in Abhängigkeit von der Zahl M weiterverwendeter Dimensionen mit und ohne Hauptachsentransformation. Die Merkmalsauswahl erfolgte nach der MS-Rangordnung.

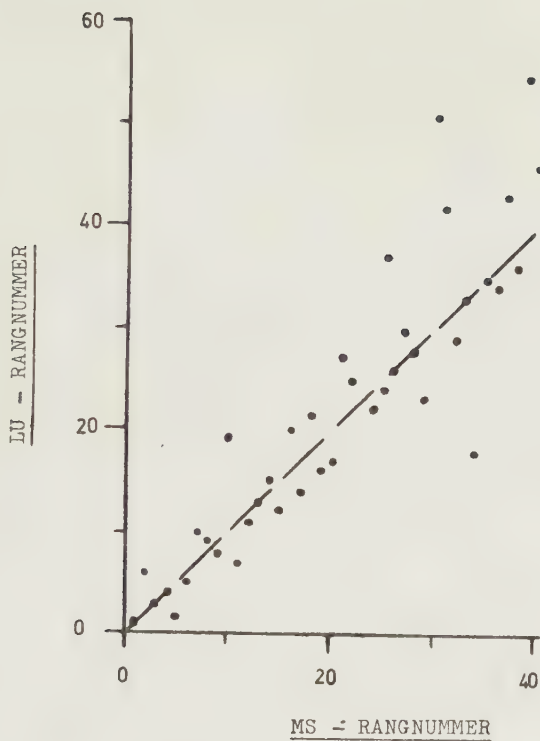


Bild 8. Zusammenhang der beiden Rangordnungen LU und MS für die Merkmale w_i (Stichprobe A0). Ein Merkmal wird durch einen Punkt repräsentiert.

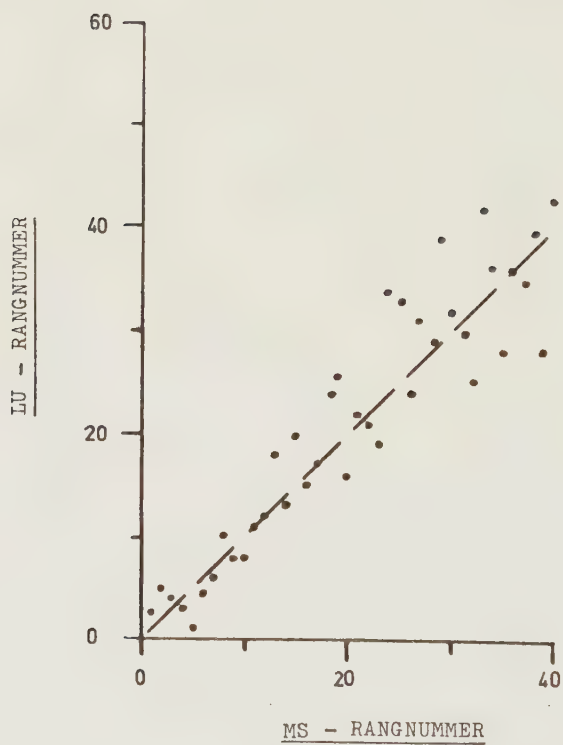


Bild 9. Zusammenhang der beiden Rangordnungen LU und MS für die Merkmale w_i (Stichprobe A1).

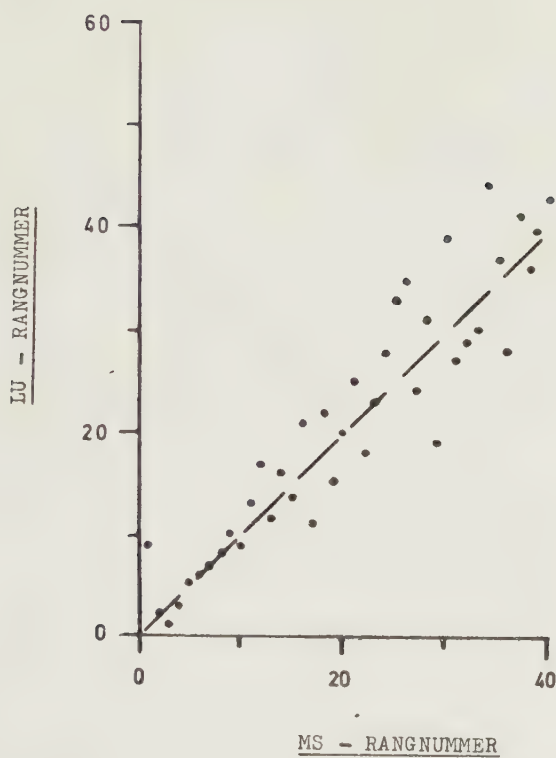


Bild 10. Zusammenhang der beiden Rangordnungen LU und MS für die Merkmale w_i (Stichprobe A2).

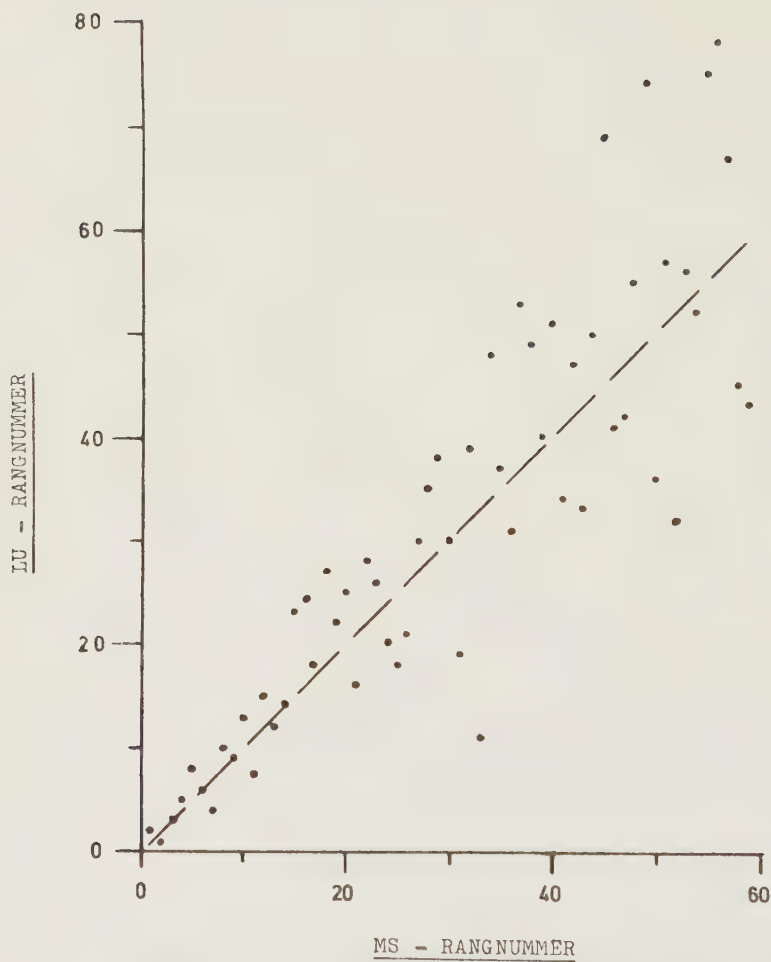
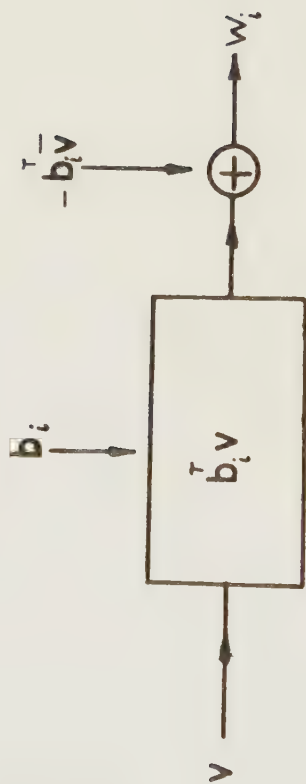


Bild 11. Zusammenhang der beiden Rangordnungen LU und MS für die Merkmale w_i (Stichprobe H0).



Bild 12. Vergleich der beiden Rangordnungen LU und MS für die Merkmale v_i (Stichprobe HO).

Eigenvektor



Merkmalsvektor

Merkmals w_i

Bild 13. Berechnung eines Merkmals w_i aus Eigenvektor b_i und Konstante ${}^T b_i v$

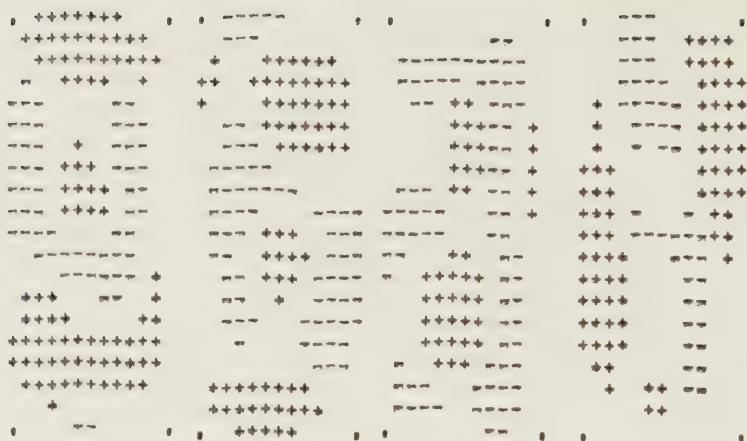
0	0	0	1	2	4	5	5	4	2	1	0
0	1	1	5	10	14	15	14	12	8	5	2
0	1	3	7	11	12	10	8	10	11	7	3
-1	0	3	5	5	-1	-5	-6	1	8	7	4
0	2	4	5	1	-5	-10	-13	-5	5	8	4
2	5	7	5	0	-4	-8	-13	-7	5	8	5
3	6	6	4	1	0	-6	-11	-5	0	9	5
4	5	4	3	4	4	-5	-9	-2	8	9	5
4	4	1	1	4	2	-8	-9	1	9	9	4
4	4	0	-3	-3	-8	-13	-10	1	9	8	4
6	6	0	-5	-8	-12	-15	-10	3	10	8	3
8	10	5	-2	-7	-12	-15	-6	9	13	8	3
9	13	10	2	-5	-10	-10	1	14	13	6	2
10	15	14	7	0	-4	-2	8	15	11	4	1
7	13	15	12	7	3	5	10	12	6	2	0
3	7	11	12	9	5	4	5	4	2	0	0
1	2	3	4	3	0	-2	-1	0	0	0	0
0	0	-1	-2	-3	-3	-1	0	0	0	0	0

EIGENVEKT,NR. 1

0	0	-2	-3	-6	-7	-4	-1	2	2	1	0
-1	-3	-6	-11	-13	-11	-8	-6	-3	1	1	1
-3	-7	-12	-12	-7	-2	-1	-6	-9	-4	0	1
-4	-9	-10	-3	6	10	6	-3	-11	-8	-2	1
-3	-5	-1	8	13	11	5	-3	-11	-9	-2	1
-1	2	9	16	14	8	1	-6	-11	-6	1	1
1	7	16	17	10	2	-3	-7	-7	1	3	2
3	12	19	13	4	-2	-6	-5	1	7	5	2
5	15	20	12	1	-6	-8	-3	8	11	7	2
6	13	18	10	-1	-8	-7	-1	9	12	8	3
5	11	11	3	-5	-7	-3	3	6	9	7	2
3	5	3	-5	-10	-7	3	8	5	6	3	0
0	0	-3	-8	-11	-3	9	11	5	2	0	-1
-3	-6	-7	-9	-8	0	9	8	0	-4	-3	-1
-4	-9	-11	-11	-9	-1	5	1	-7	-8	-4	-2
-2	-6	-11	-12	-9	-2	3	-1	-6	-4	-2	-1
0	-2	-4	-6	-4	1	5	3	0	-1	0	0
0	-1	-2	-2	-1	2	3	2	1	0	0	0

EIGENVEKT,NR. 2

Bild 14. Eigenvektoren der Kovarianzmatrix der Stichprobe H0 in Rasterbilddarstellung. Jede Komponente der Vektoren wurde mit 100 multipliziert, gerundet, und dann in das zugehörige Rasterbildelement gedruckt.

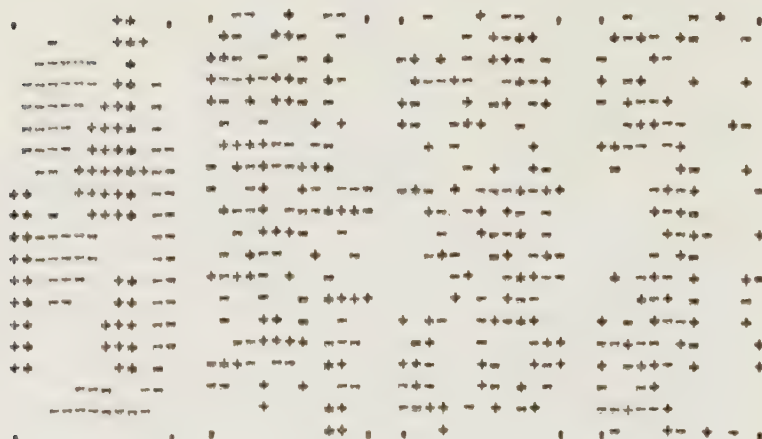


NR.1

NR.2

NR.3

NR. 4



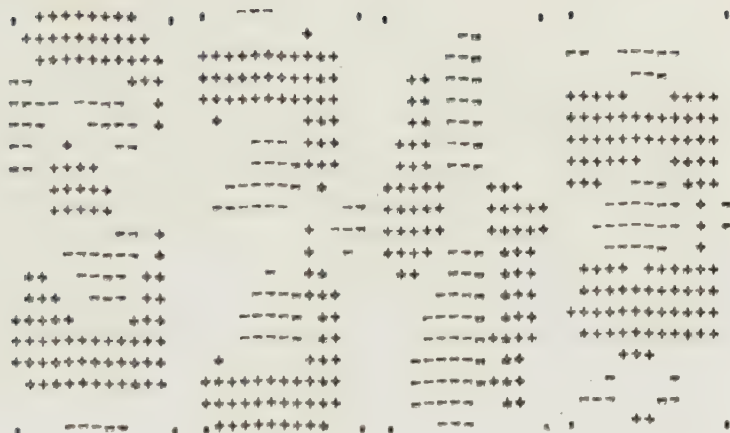
NR.5

NR.100

NR.150

NR.200

Bild 15. Eigenvektoren der Kovarianzmatrix der Stichprobe A0
in Rasterbilddarstellung.

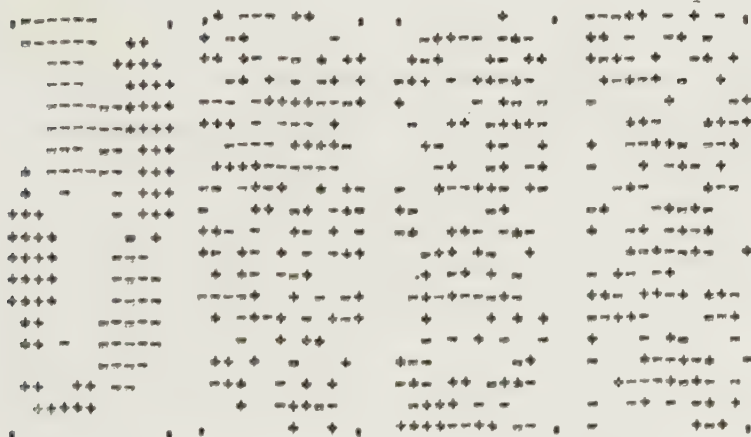


NR.1

NR.2

NR.3

NR.4



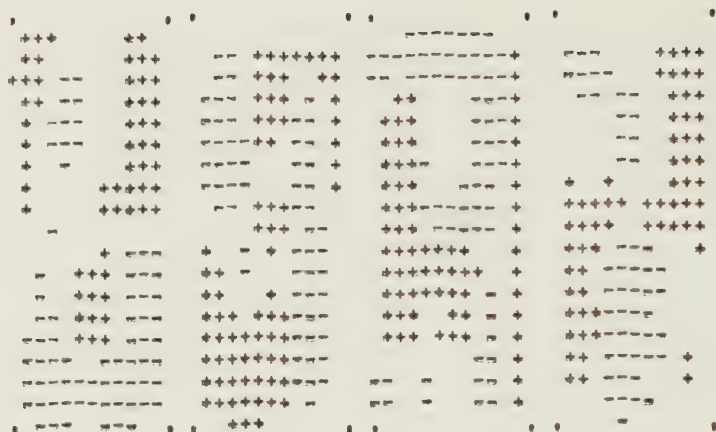
NR.5

NR.100

NR.150

NR.200

Bild 16. Eigenvektoren der Kovarianzmatrix der Stichprobe A1 in Rasterbilddarstellung.

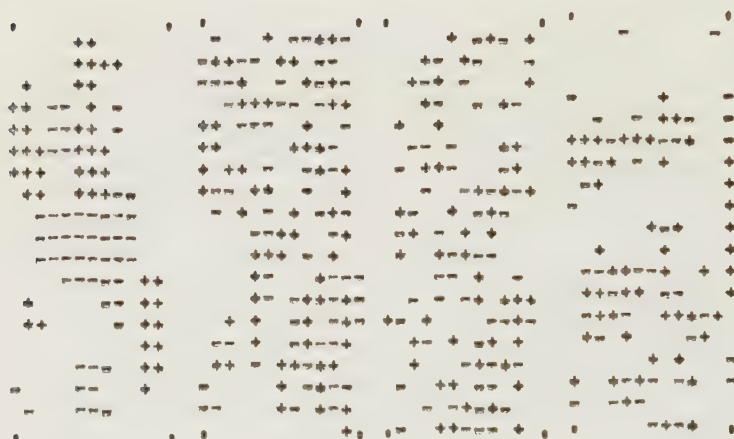


NR.1

NR.2

NR.3

NR.4



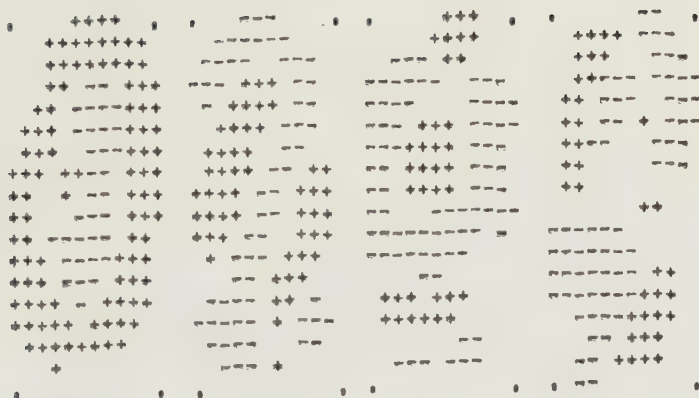
NR.5

NR.100

NR.150

NR.200

Bild 17. Eigenvektoren der Kovarianzmatrix der Stichprobe A2
in Rasterbilddarstellung.

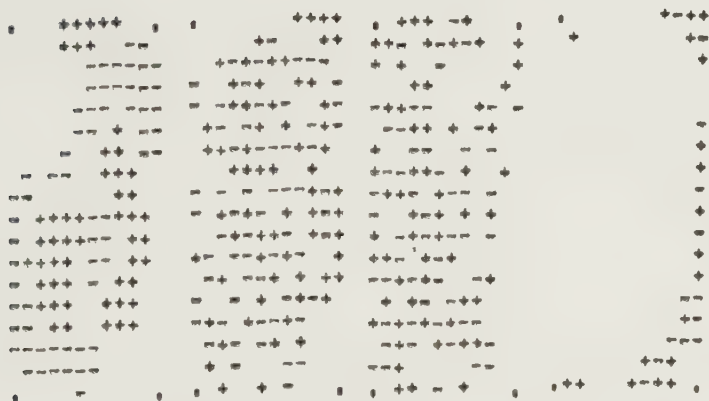


NR.1

NR.2

NR.3

NR.4



NR.5

NR.100

NR.150

NR.200

BiId 18. Eigenvektoren der Kovarianzmatrix der Stichprobe H0
in Rasterbilddarstellung.

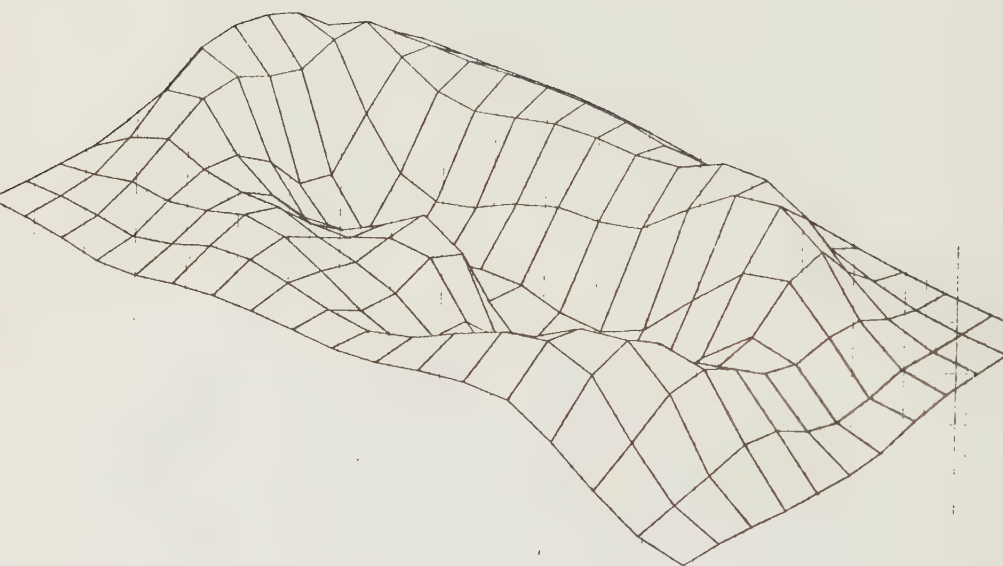


Bild 19. Perspektivische Darstellung des 1. Eigenvektors der Kovarianzmatrix der Stichprobe H_0 .

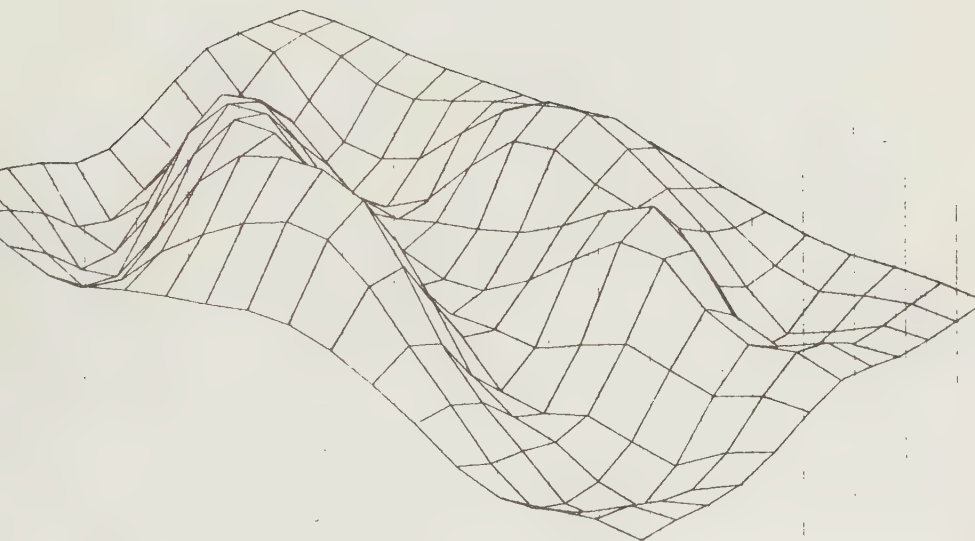


Bild 20. Perspektivische Darstellung des 2. Eigenvektors der Kovarianzmatrix der Stichprobe H_0 .

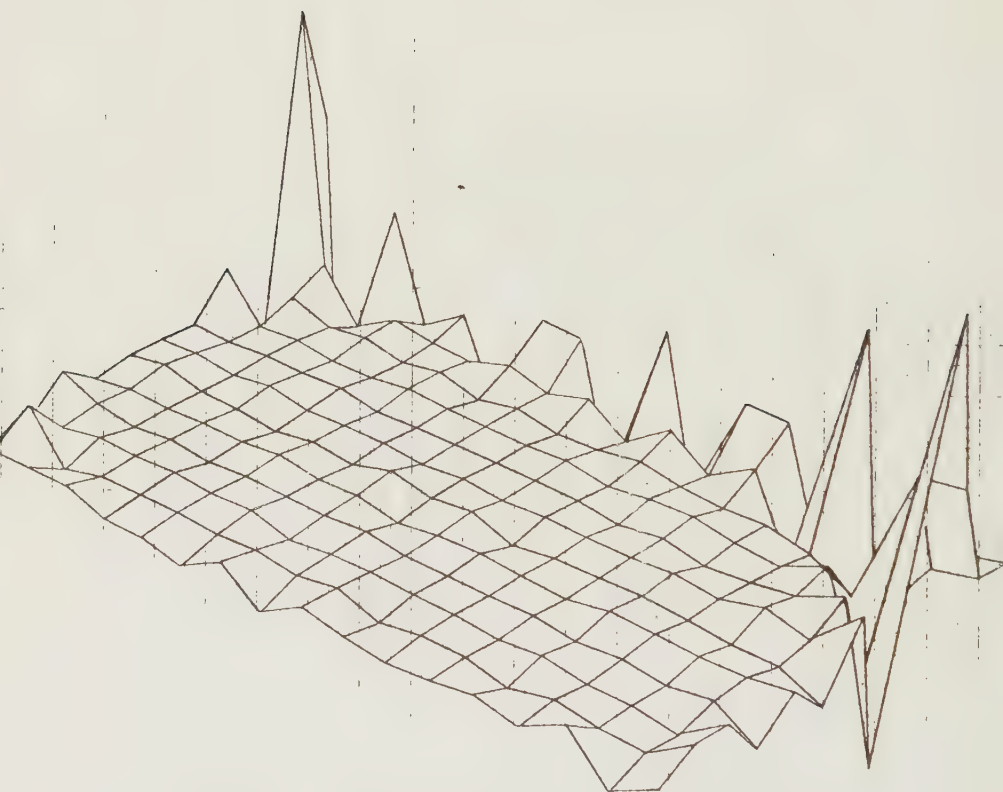


Bild 21. Perspektivische Darstellung des 200. Eigenvektors
der Kovarianzmatrix der Stichprobe H0.

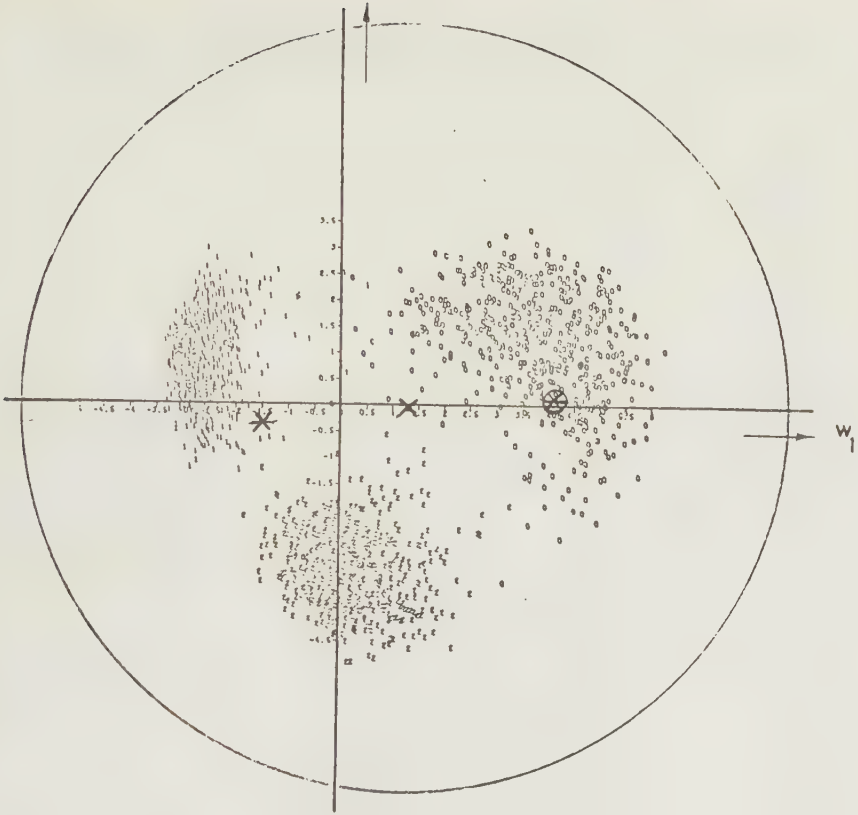


Bild 22. Darstellung des N-dimensionalen Merkmalsraumes in der $w_1 w_2$ Ebene. Jedes Zeichen wird im Merkmalsraum durch einen Punkt repräsentiert. Die zu den Zeichenklassen "1", "2" und "0" gehörigen Punkte wurden auf diese Ebene projiziert und markiert. Der eingezeichnete Kreis begrenzt die Fläche, in der überhaupt Zeichen vorkommen können. Mit eingezeichnet sind die Bilder von $v_0 = (0, 0, 0, \dots 0)^T$ (*), $v_{0,5} = (0,5, 0,5, \dots 0,5)^T$ (x), $v_1 = (1, 1, 1 \dots 1)^T$ (⊗).

WERTEVERTEILUNG DER VEKTORKOMPONENTEN

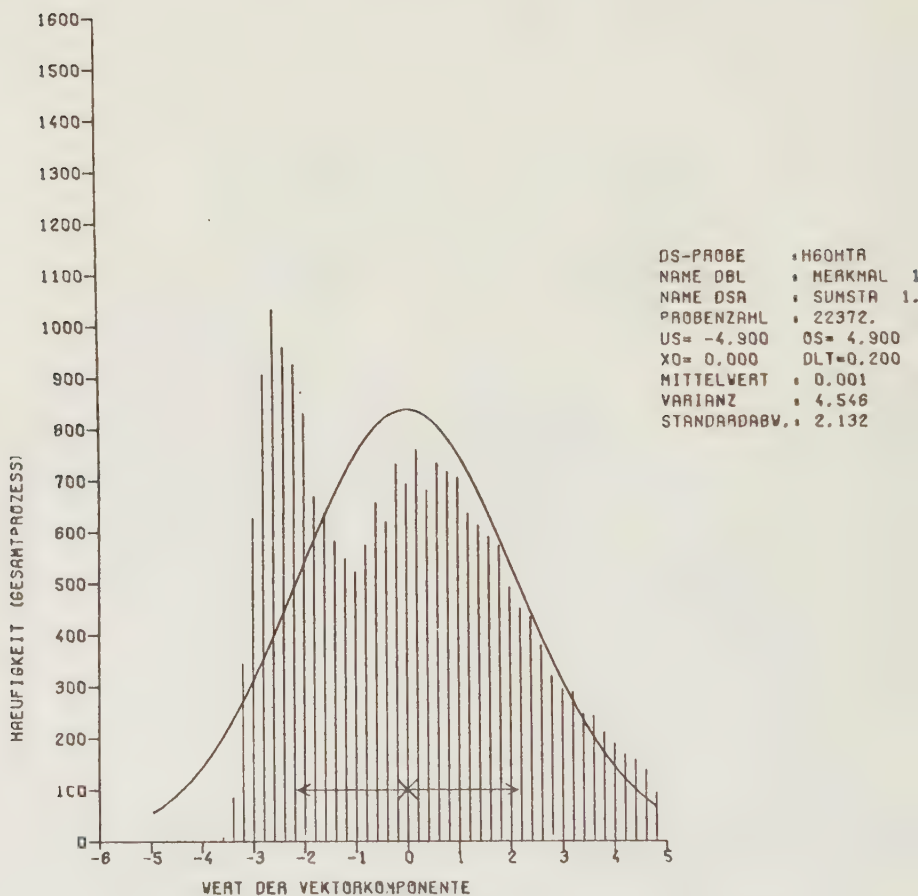


Bild 23. Verteilung von w_1 ; Stichprobe H0.

WERTEVERTEILUNG DER VEKTORKOMPONENTEN

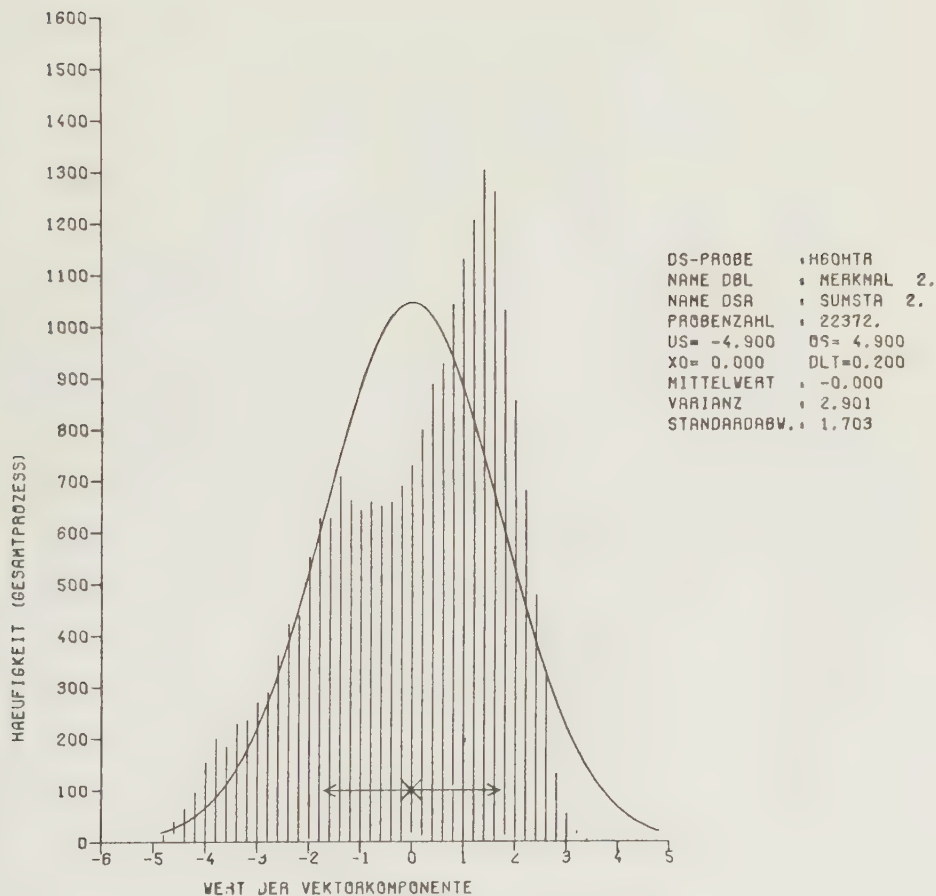


Bild 24. Verteilung von w_2 ; Stichprobe H0.

WERTEVERTEILUNG DER VEKTORKOMPONENTEN

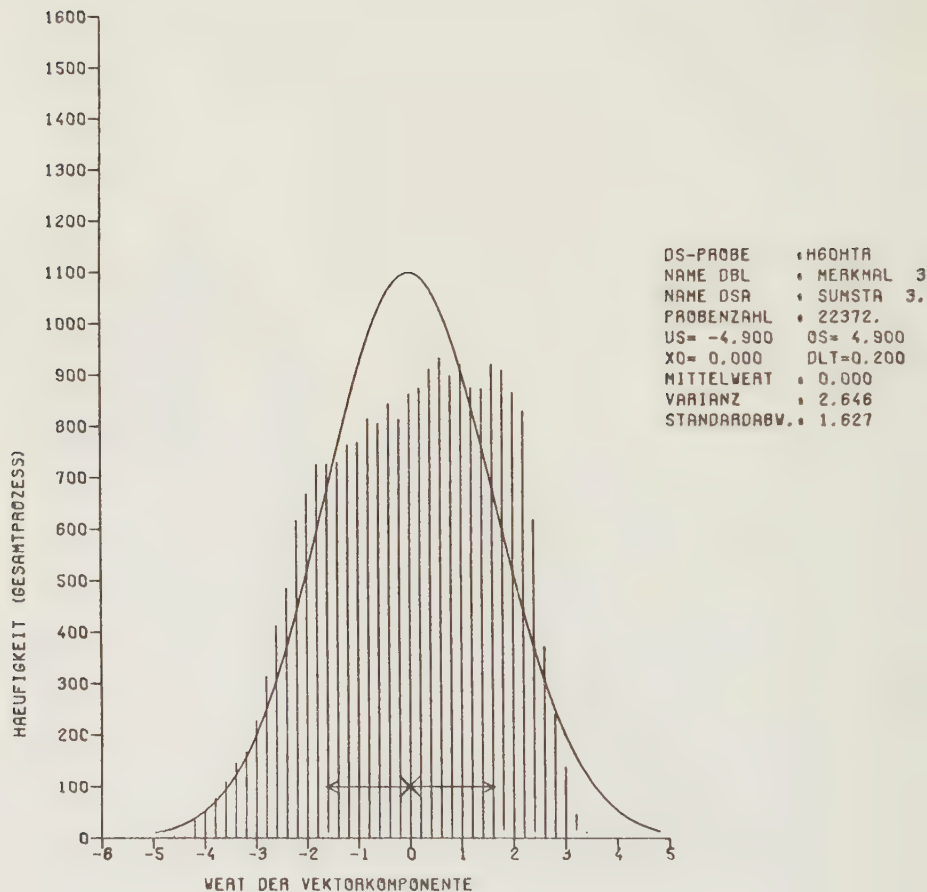
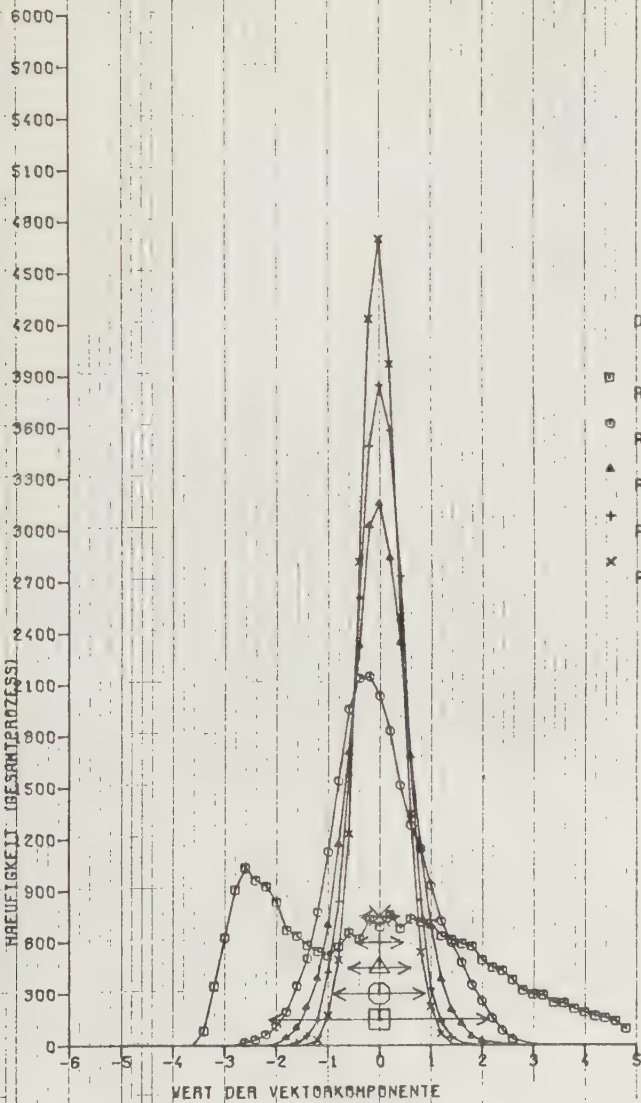


Bild 25. Verteilung von w_3 ; Stichprobe H0.

WERTEVERTEILUNG DER VEKTORKOMPONENTEN



DS-PROBE : H6QHTA

```

W MEARKMAL 1. SUMSTA 1.
  ANZ=22372, MV=0.001 VAR.=4.546
U MEARKMAL 10. SUMSTA 10.
  ANZ=22372, MV=0.000 VAR= 0.784
A MEARKMAL 20. SUMSTA 20.
  ANZ=22372, MV=-0.000 VAR= 0.362
+ MEARKMAL 30. SUMSTA 30.
  ANZ=22372, MV=-0.000 VAR= 0.213
X MEARKMAL 40. SUMSTA 40.
  ANZ=22372, MV=0.000 VAR= 0.146

```

Bild 26. Verteilungen von $w_1, w_{10}, w_{20}, w_{30}, w_{40}$;
Stichprobe H0.

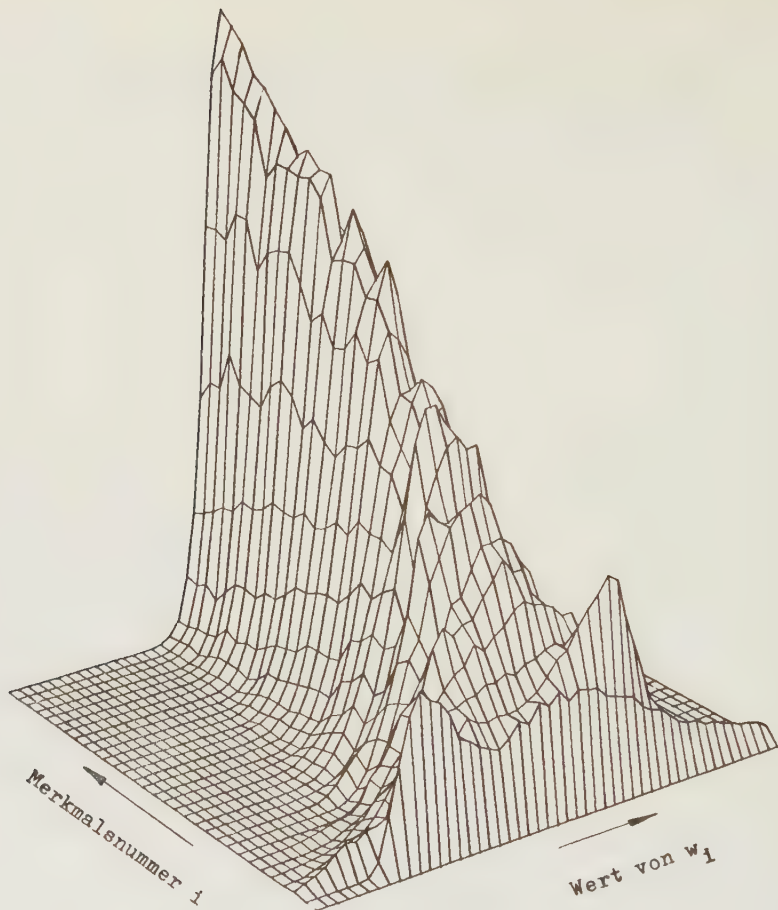


Bild 27. Perspektivische Darstellung der Verteilungen der Merkmale w_1 bis w_{30} ; Stichprobe H0.

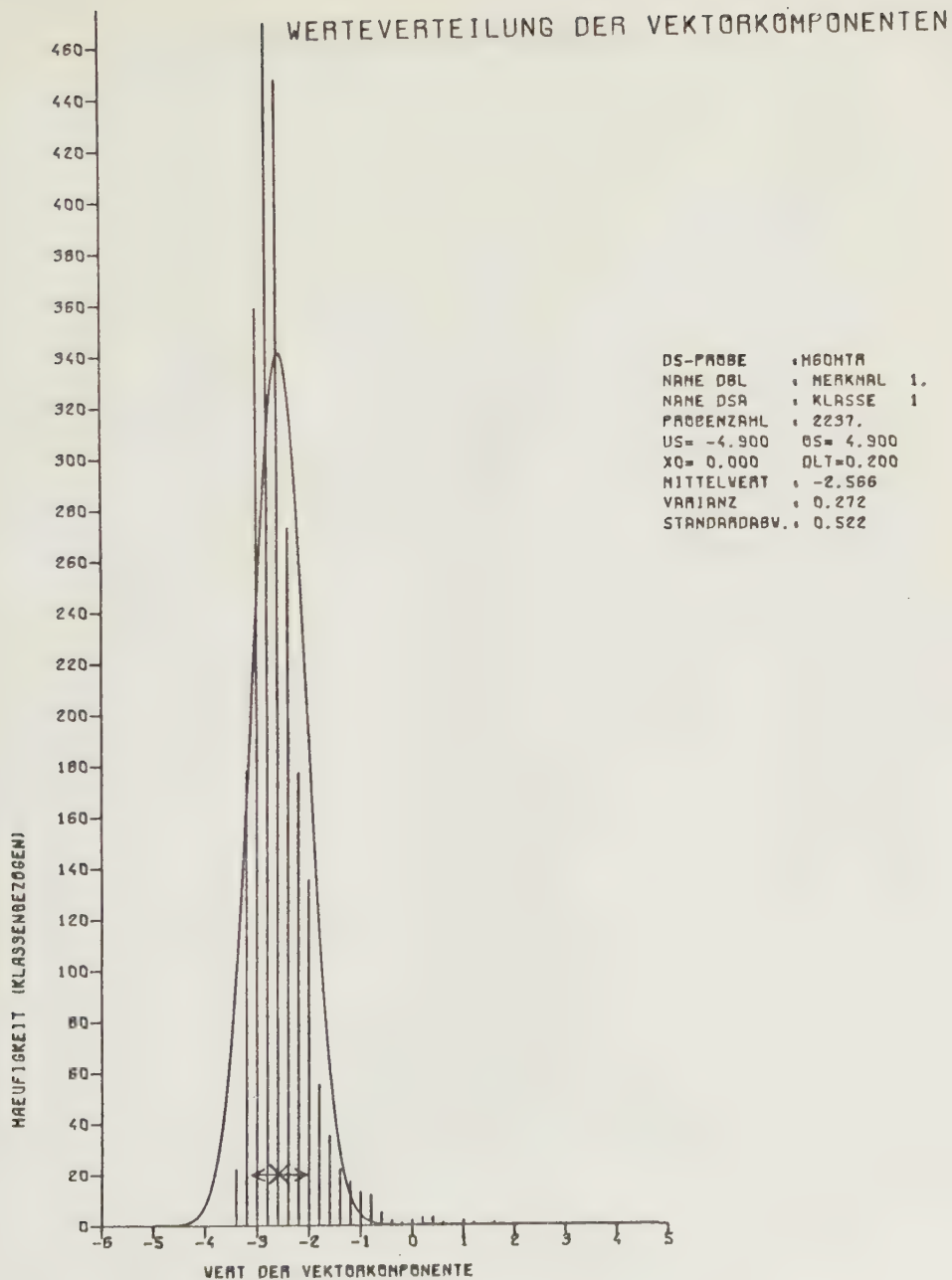
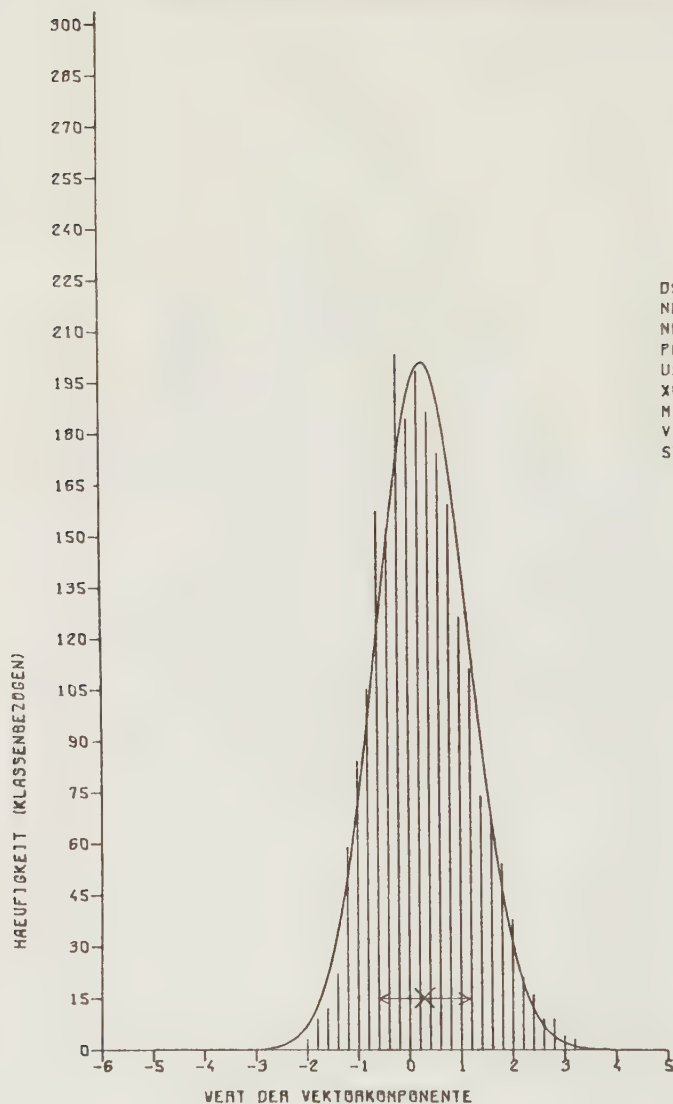


Bild 28. Verteilung von $w_{1/k=1}$; Stichprobe H0.

WERTEVERTEILUNG DER VEKTORKOMPONENTEN



DS-PROBE : H6QHTA
 NAME DEL : MEAKMAL 1.
 NAME DSA : KLASSE 2
 PROBENZAHL : 2239.
 US= -4.900 OS= 4.900
 X0= 0.000 DLT=0.200
 MITTELWERT : 0.280
 VARIANZ : 0.792
 STANDARDABW. : 0.890

Bild 29. Verteilung von $w_{1/k=2}$; Stichprobe H0.

WERTEVERTEILUNG DER VEKTORKOMPONENTEN

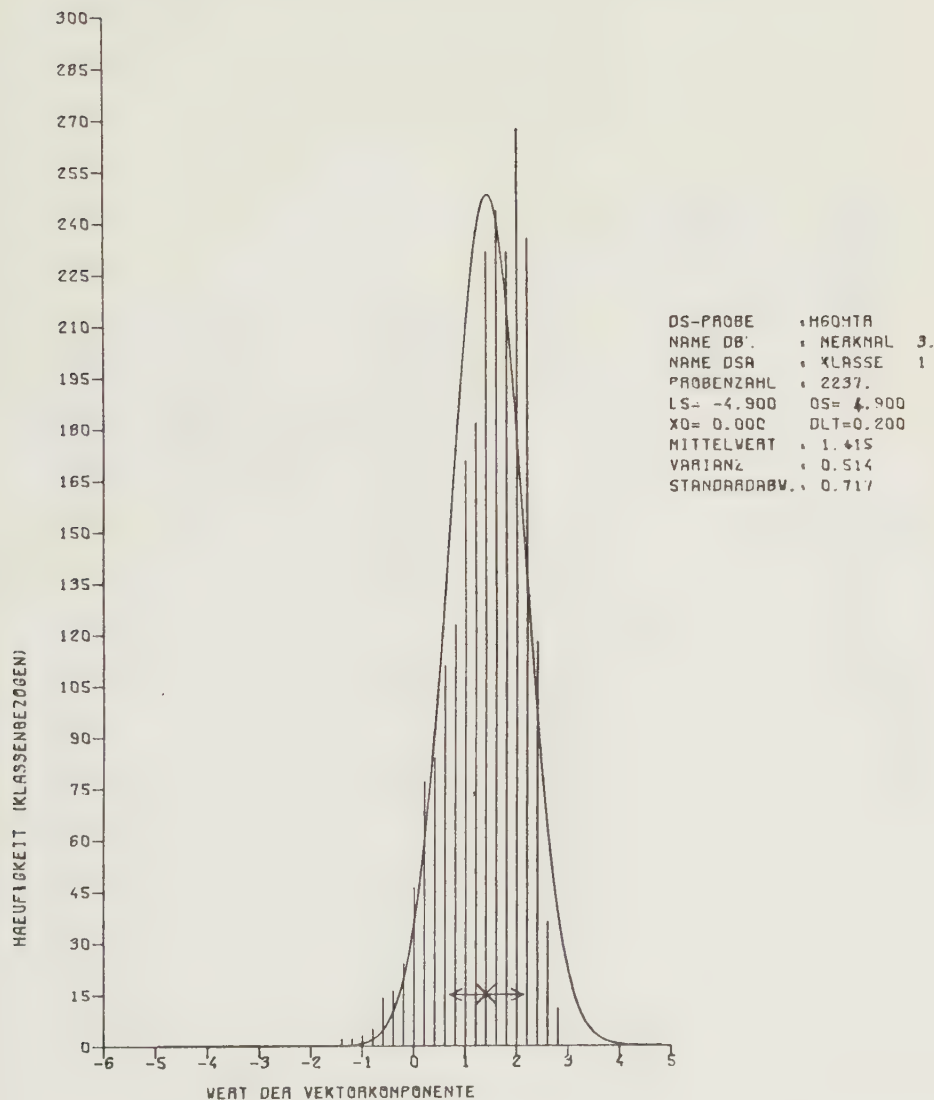


Bild 30. Verteilung von $w_{3/k=1}$; Stichprobe H0.

WERTEVERTEILUNG DER VEKTORKOMPONENTEN

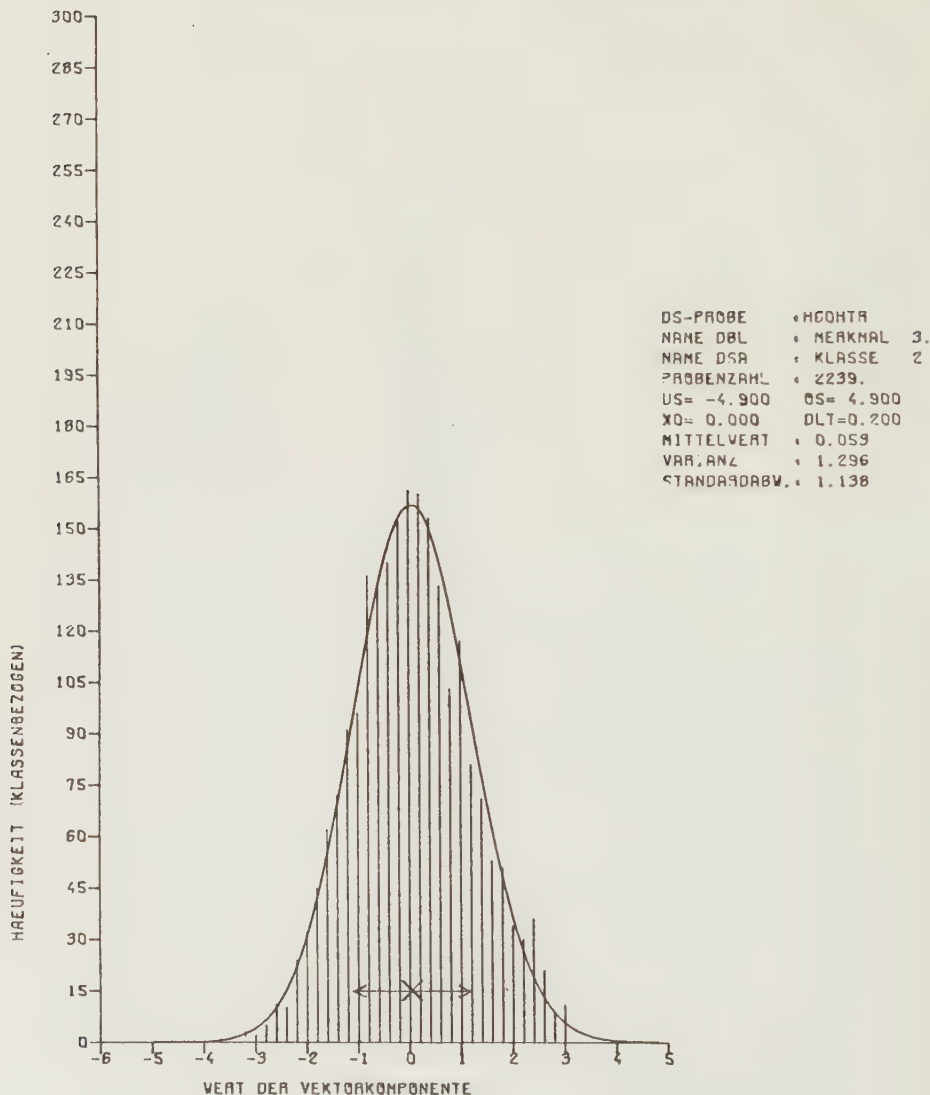
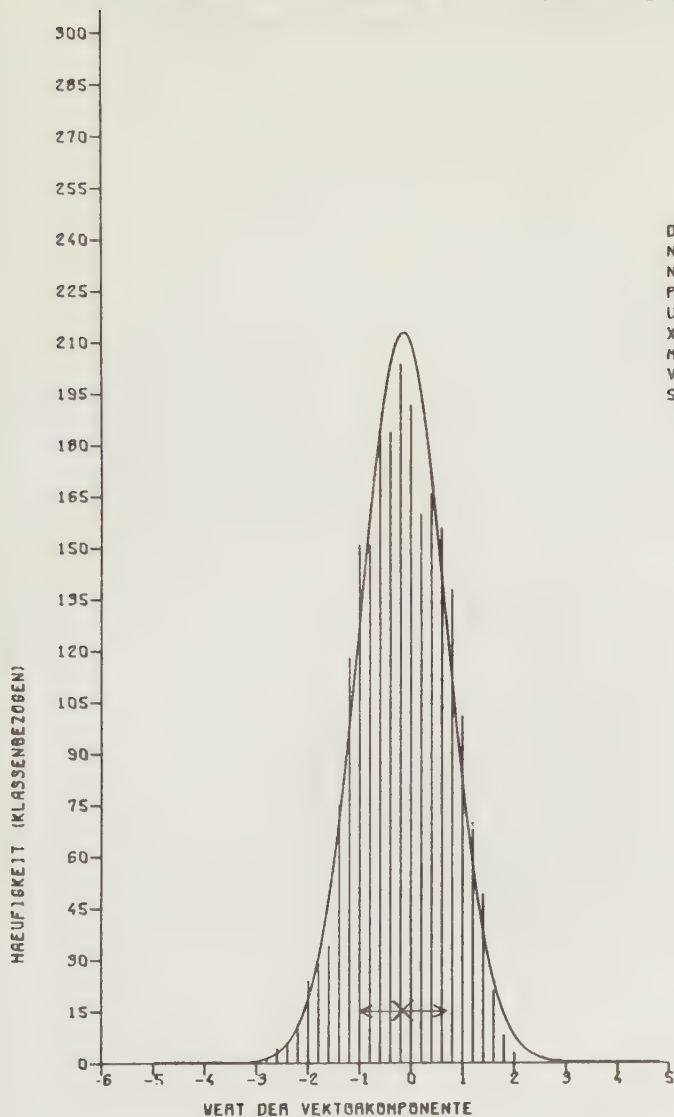


Bild 31. Verteilung von $w_{3/k=2}$; Stichprobe H0.

WERTEVERTEILUNG DER VEKTORKOMPONENTEN



DS-PROBE : H60HTA
 NAME DBL : MERKMAL 13.
 NAME DSA : KLASSE 1
 PROBENZAHL : 2237.
 US= -4.900 OS= 4.900
 XO= 0.000 DLT=0.200
 MITTELWERT : -0.153
 VARIANZ : 0.700
 STANDARDABW. : 0.837

Bild 32. Verteilung von $w_{13/k=1}$; Stichprobe H0.

WERTEVERTEILUNG DER VEKTORKOMPONENTEN

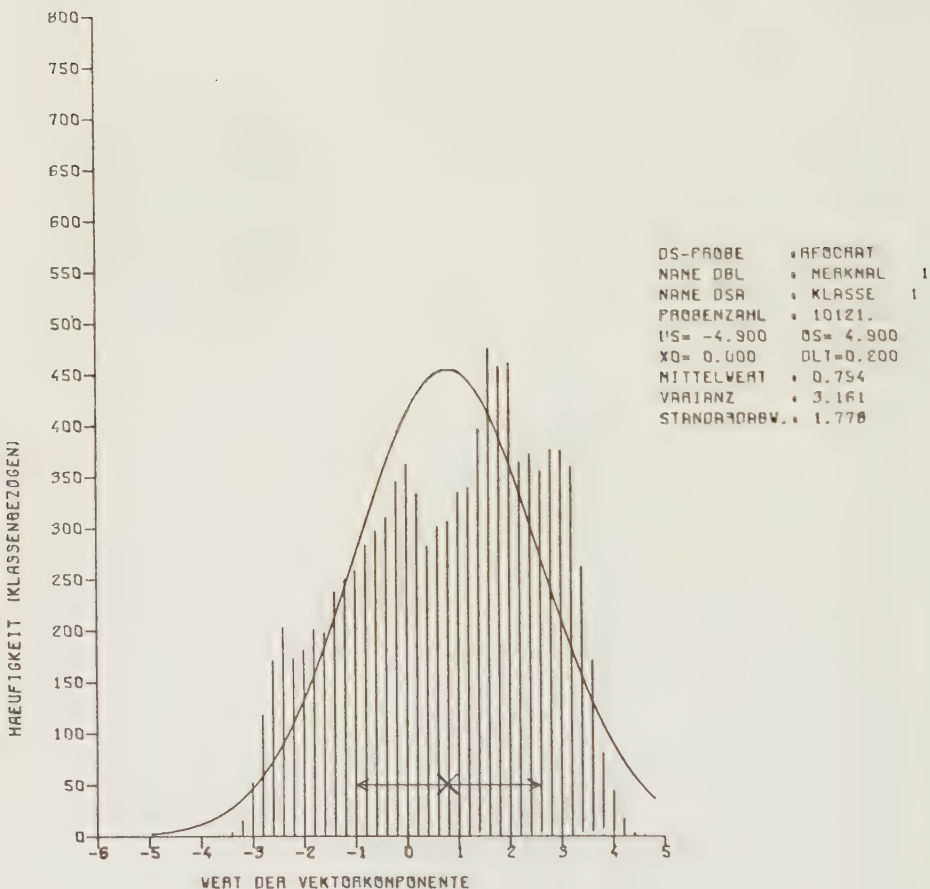


Bild 33. Verteilung von $w_{1/k=1}$; Stichprobe A0.

WERTEVERTEILUNG DER VEKTORKOMPONENTEN

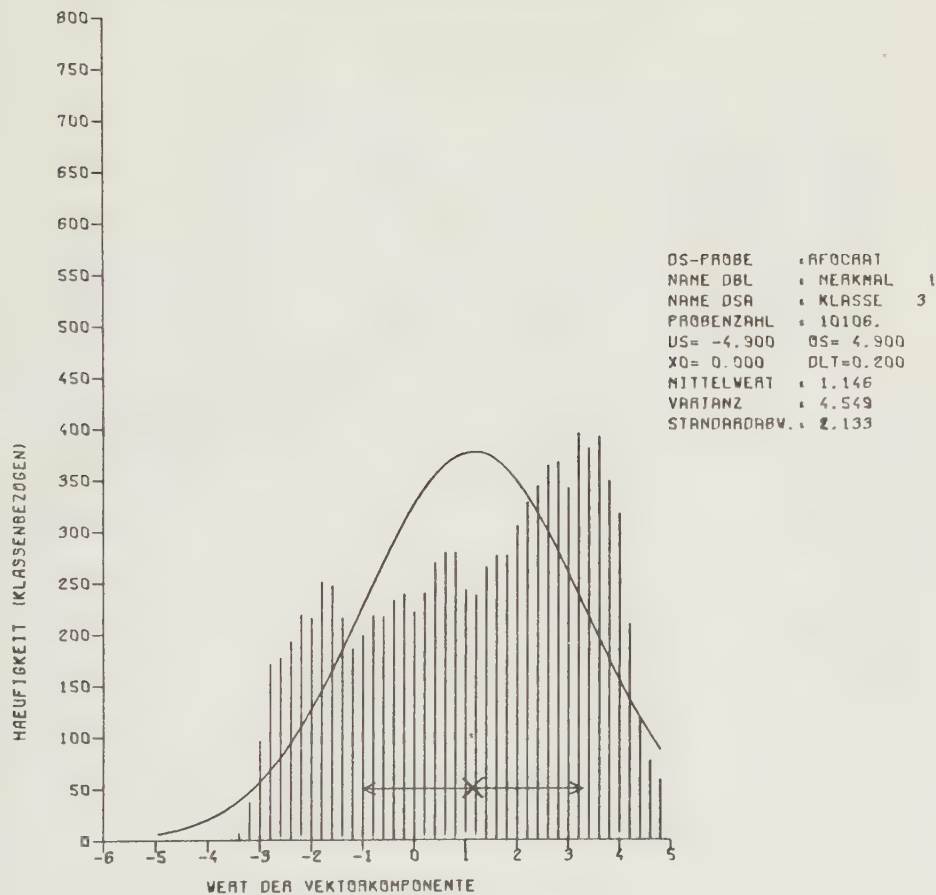
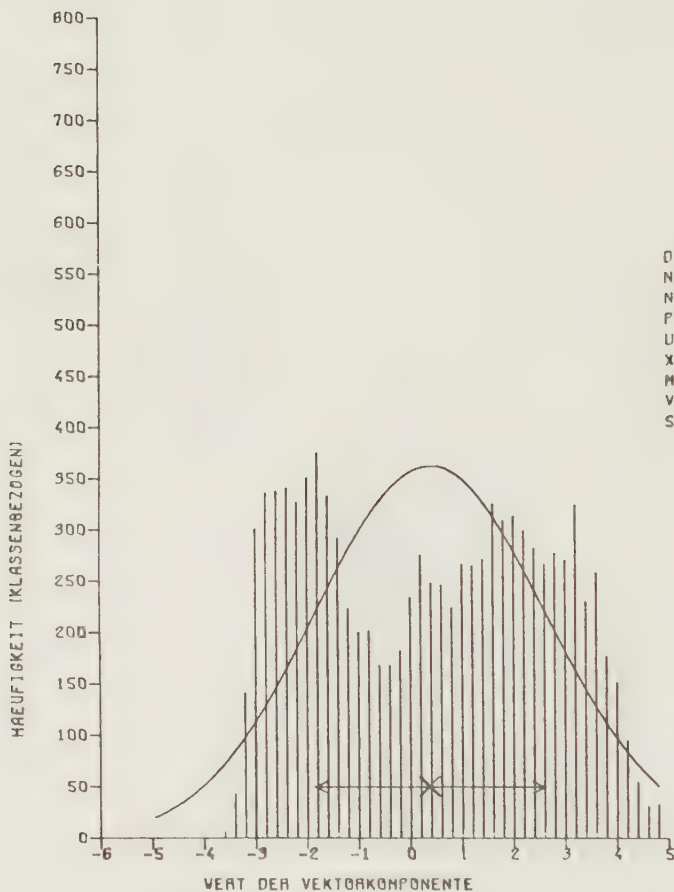


Bild 34. Verteilung von $w_{1/k=3}$; Stichprobe A0.

WERTEVERTEILUNG DER VEKTORKOMPONENTEN



DS-PROBE : AF00ART
 NAME DBL : MERKMAL 1
 NAME DSA : KLASSE 4
 PROBENZAHL : 10082.
 US= -4.900 BS= 4.900
 X0= 0.000 DLT=0.200
 MITTELWERT : 0.983
 VARIANZ : 4.939
 STANDARDABW. : 2.222

Bild 35. Verteilung von $w_{1/k=4}$; Stichprobe A0.

WERTEVERTEILUNG DER VEKTORKOMPONENTEN

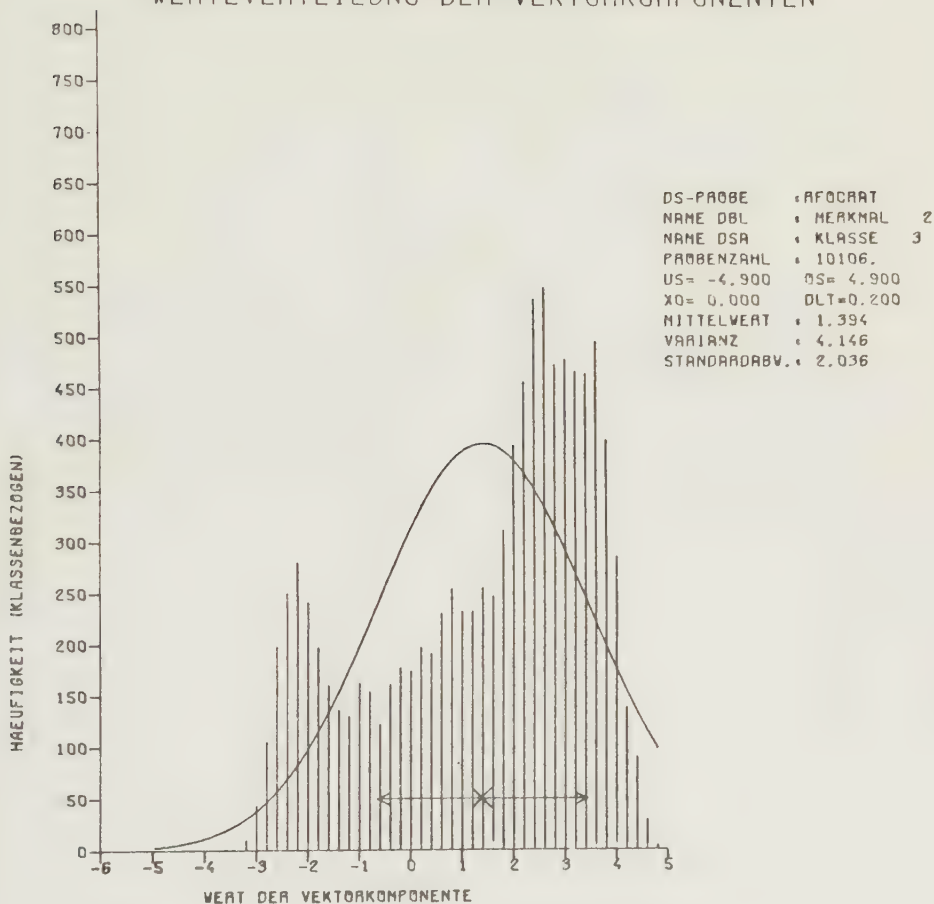


Bild 36. Verteilung von $w_{2/k=3}$; Stichprobe A0.

WERTEVERTEILUNG DER VEKTORKOMPONENTEN

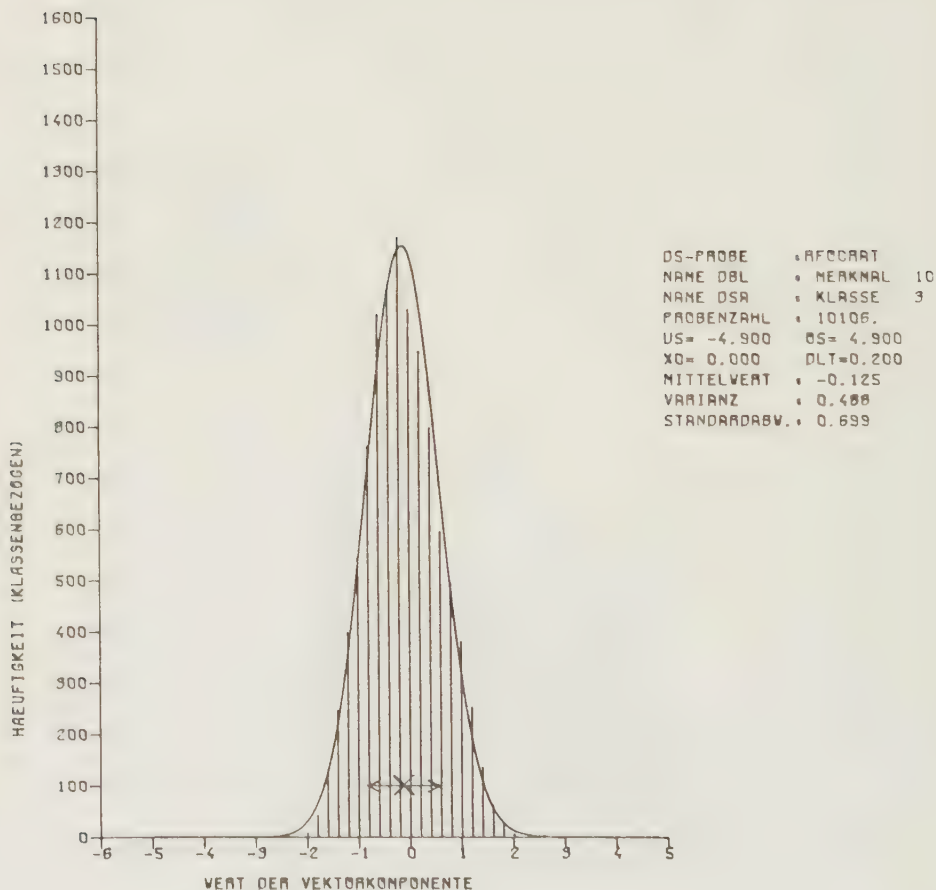


Bild 37. Verteilung von $w_{10/k=3}$; Stichprobe A0.

WERTEVERTEILUNG DER VEKTORKOMPONENTEN

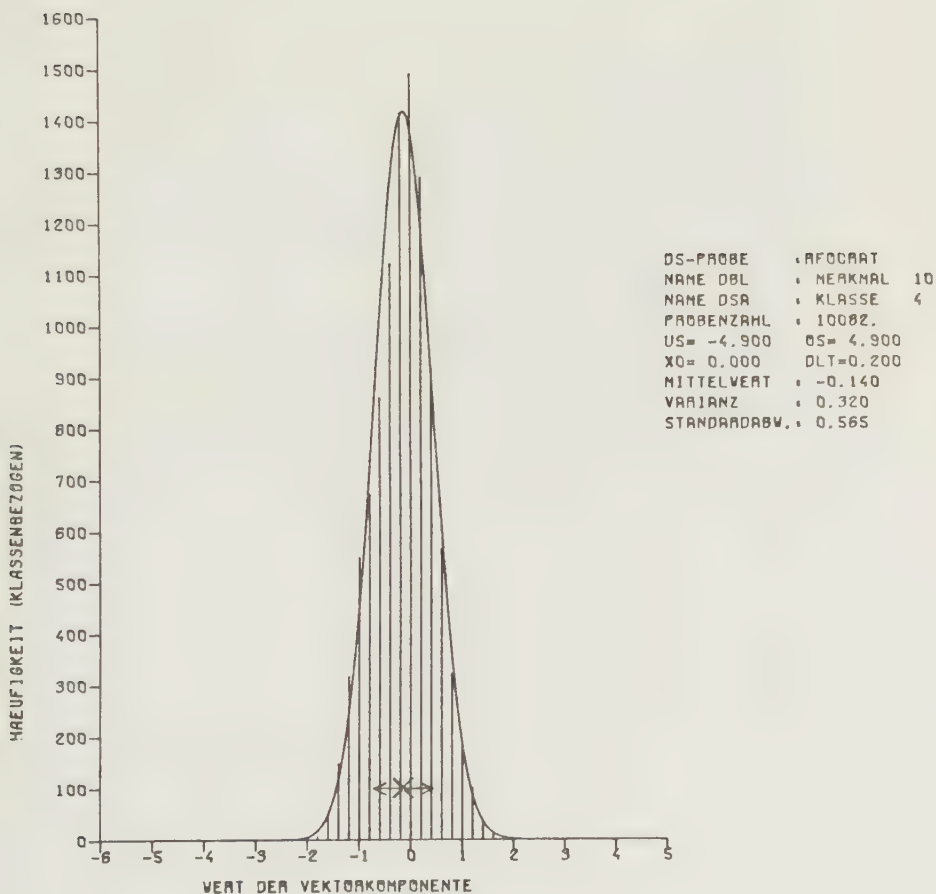
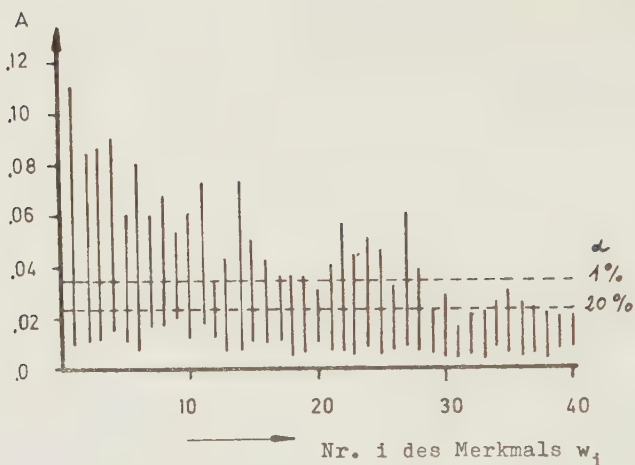
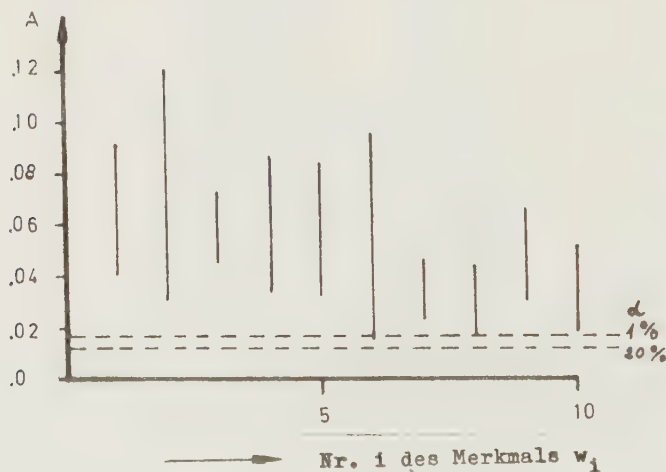


Bild 38. Verteilung von $w_{10/k=4}$; Stichprobe A0.



2 200 Repräsentanten pro Klasse; Handblockschrift



10 000 Repräsentanten pro Klasse; OCR-A

Bild 39. Ergebnisse des Kolmogoroff-Smirnow-Tests für die Verteilung der klassenspezifischen Merkmale $w_{i/k}$ der Handblockschrift-Stichprobe H0 und der OCR-A Stichprobe AO. Modellkurve: Normalverteilung.

REQUIVALENTE ELLIPSEN DK=KONST.

v_2

DS-PROBE H60HTA
 ++ MESSWERTE IN EBENE v_1, v_2 .
 KENNUNG 0 PROBENZAHL 2230.
 $v_1 v_2$ KENNUNG 0
 MITTELPUNKT 3.57 1.12
 DK =2.00 $A=1.94$ $B=1.44$

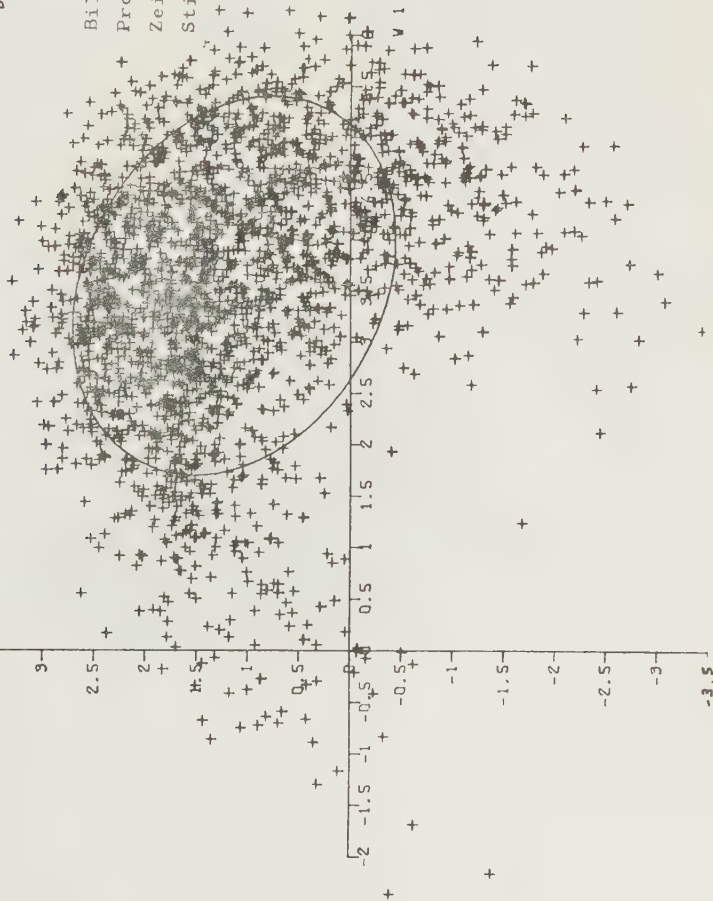
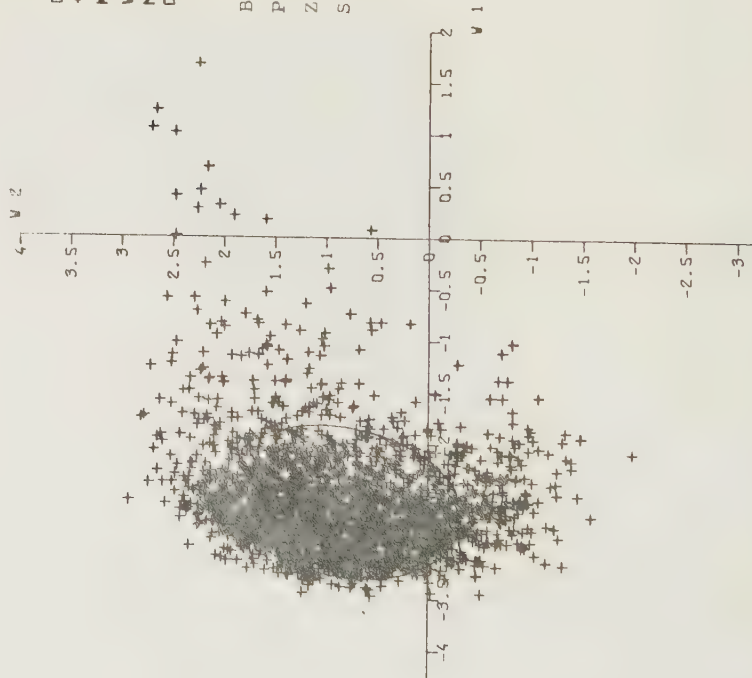


Bild 40.

Projektion auf die Ebene $w_1 w_2$
 Zeichen mit Kennung "0"
 Stichprobe H0.

REQUIVALENTE ELLIPSEN DK=KONST.

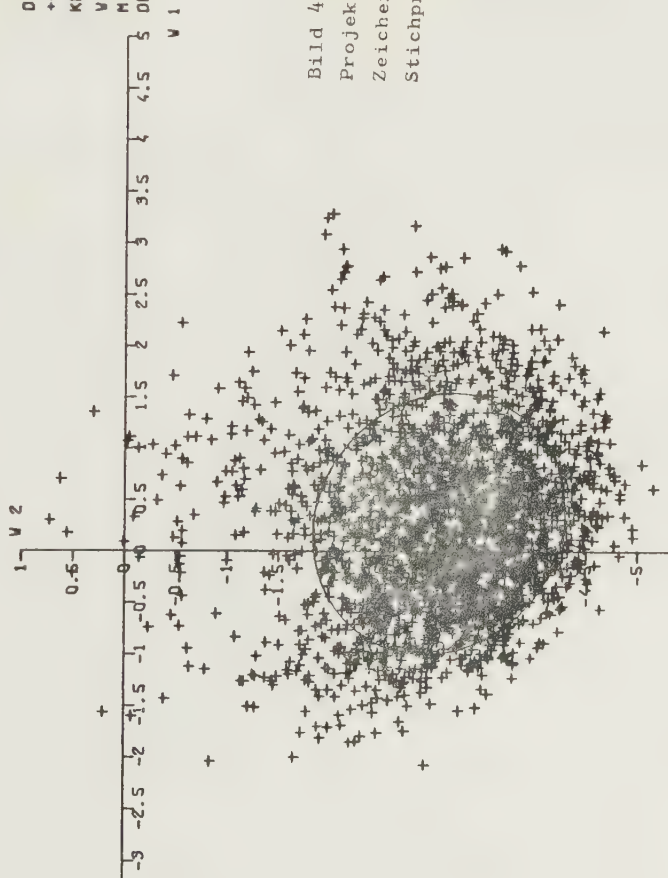


DS-PROBE HÖCHST
 +++ MESSWERTE IN EBENE w_1 , w_2 .
 KENNUNG 1 PROBENZAHL, 2237.
 w_1 w_2 KENNUNG 1
 MITTELPUNKT, -2.57 0.84
 DK = 2.00 $A = 1.17$ $B = 0.72$

Bild 41.

Projektion auf die Ebene $w_1 w_2$
 Zeichen mit Kennung "1"
 Stichprobe H0.

REQUIVALENTE ELLIPSEN $DK=KONST.$

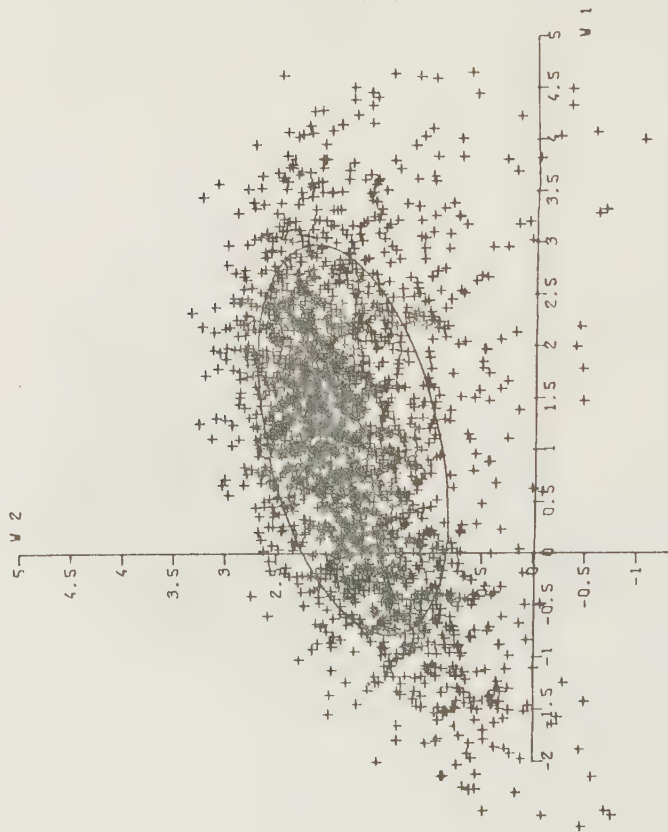


DS-PROBE M60HTA
 +++ MESSWEITE IN EBENE w_1, w_2 .
 KENNUNG 2 PROBENZAHL: 2239.
 $w_1 w_2$ KENNUNG 2
 MITTELPUNKT: 0.29 -3.08
 $DK = 2.00$ $A = 1.31$ $B = 1.17$

Bild 42.

Projektion auf die Ebene $w_1 w_2$
 Zeichen mit Kennung "3"
 Stichprobe H_0 .

REQUIVALENTE ELLIPSEN $DK = \text{KONST.}$



DS-PROBE H60HTA
 ++ MESSWERTE IN EBENE v_1, v_2 .
 KENNUNG 6 PROBENZAHL: 2236.
 $v_1 w_2$ KENNUNG 6
 MITTELPUNKT: 1.13 1.76
 $DK = 2.00$ $A = 1.91$ $B = 0.83$

Bild 43.

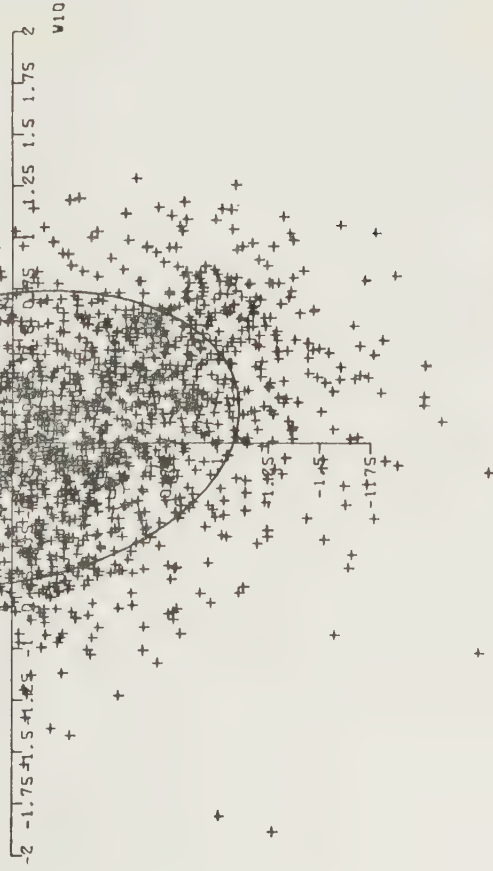
Projektion auf die Ebene $w_1 w_2$
 Zeichen mit Kennung "6"
 Stichprobe H0.

REQUIVALENTE ELLIPSEN $DK=KÖNST.$

DS-PROBE H6QHTA
++ MESSWEITE IN EBENE W_{10} W_{20} .
KENNUNG 7 PROBENZAHL 2238.
 $W_{10}W_{20}$ KENNUNG 7
MITTELPUNKT, Q.02 Q.03
DK =2.00 A =1.14 B =0.70

Bild 44.

Projektion auf die Ebene $W_{10}W_{20}$
Zeichen mit Kennung "7"
Stichprobe 110.



3.5 v 2 REQUIVALENTE ELLIPSEN DK=KÖNSTE.

DS-PROBE H60HTA

V 1V 2 KENNUNG 1
 MITTELPUNKT, -2.57 0.84
 DK =2.00 A =1.17 B =0.72
 V 1V 2 KENNUNG 2
 MITTELPUNKT, 0.29 -3.08
 DK =2.00 A =1.31 B =1.17
 V 1V 2 KENNUNG 3
 MITTELPUNKT, 0.46 -1.53
 DK =2.00 A =1.47 B =0.95
 V 1V 2 KENNUNG 4
 MITTELPUNKT, -1.71 1.36
 DK =2.00 A =1.07 B =0.75
 V 1V 2 KENNUNG 5
 MITTELPUNKT, 1.73 0.34
 DK =2.00 A =1.70 B =0.31
 V 1V 2 KENNUNG 6
 MITTELPUNKT, 1.13 1.76
 DK =2.00 A =1.91 B =0.83
 V 1V 2 KENNUNG 7
 MITTELPUNKT, -1.99 -1.48
 DK =2.00 A =1.40 B =0.98
 V 1V 2 KENNUNG 8
 MITTELPUNKT, 0.92 -0.62
 DK =2.00 A =1.72 B =1.04
 V 1V 2 KENNUNG 9
 MITTELPUNKT, -1.83 1.28
 DK =2.00 A =1.32 B =0.90
 V 1V 2 KENNUNG 0
 MITTELPUNKT, 3.57 1.12
 DK =2.00 A =1.94 B =1.44

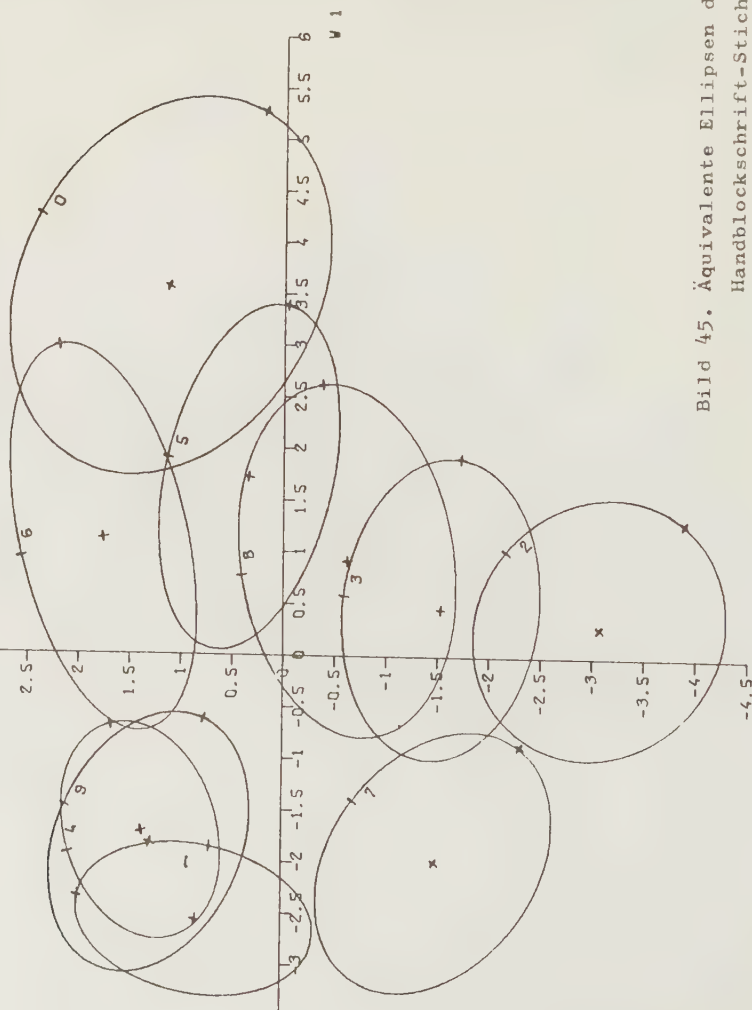


Bild 45. Äquivalente Ellipsen der 10 Klassen der

Handblockschrift-Stichprobe in der Ebene $w_1 w_2$.

REQUIVALENTE ELLIPSEN DK=KONST.

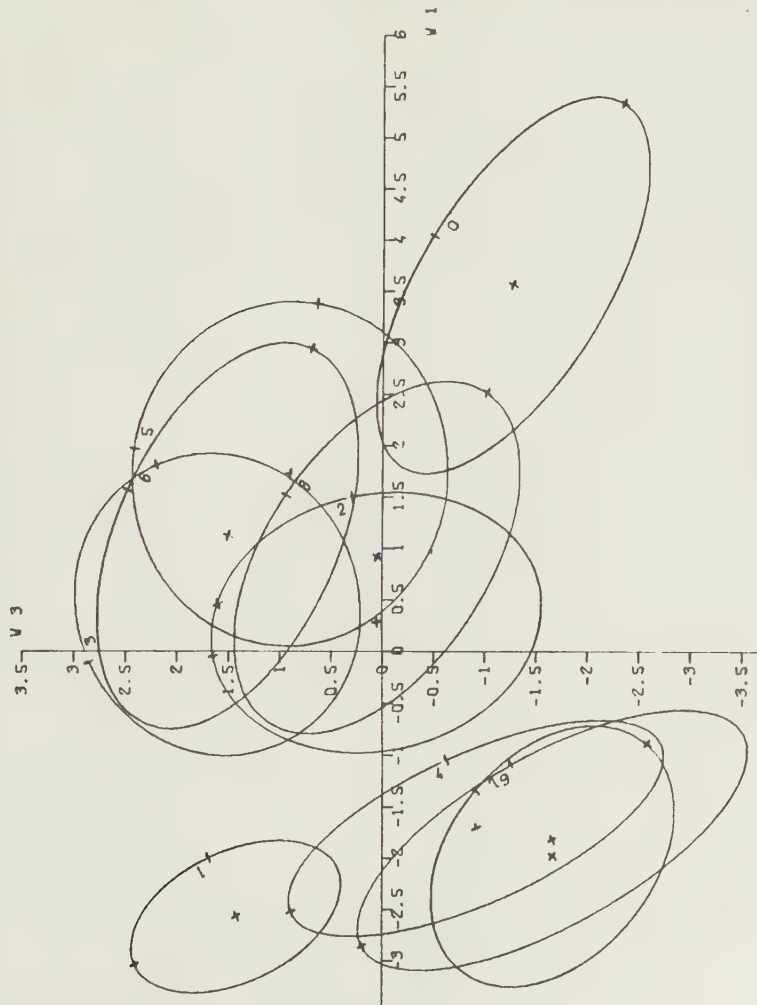


Bild 46. Äquivalente Ellipsen der 10 Klassen der Handblock-
 schrift-Stichprobe in der Ebene w_1w_3 .

DS-PROBE

H6QHTA

V 1W 3

KENNUNG 1

MITTELPUNKT, -2.57 1.41

DK =2.00 A =1.08 B =0.64

V 1W 3

KENNUNG 2

MITTELPUNKT, 0.29 0.05

DK =2.00 A =1.62 B =1.24

V 1W 3

KENNUNG 3

MITTELPUNKT, 0.46 1.60

DK =2.00 A =1.48 B =1.37

V 1W 3

KENNUNG 4

MITTELPUNKT, -1.71 -0.92

DK =2.00 A =1.98 B =0.73

V 1W 3

KENNUNG 5

MITTELPUNKT, 1.73 0.90

DK =2.00 A =1.67 B =1.54

V 1W 3

KENNUNG 6

MITTELPUNKT, 1.13 1.50

DK =2.00 A =1.99 B =1.07

V 1W 3

KENNUNG 7

MITTELPUNKT, -1.99 -1.67

DK =2.00 A =1.43 B =0.99

V 1W 3

KENNUNG 8

MITTELPUNKT, 0.92 0.05

DK =2.00 A =1.92 B =1.08

V 1W 3

KENNUNG 9

MITTELPUNKT, -1.83 -1.67

DK =2.00 A =2.12 B =0.84

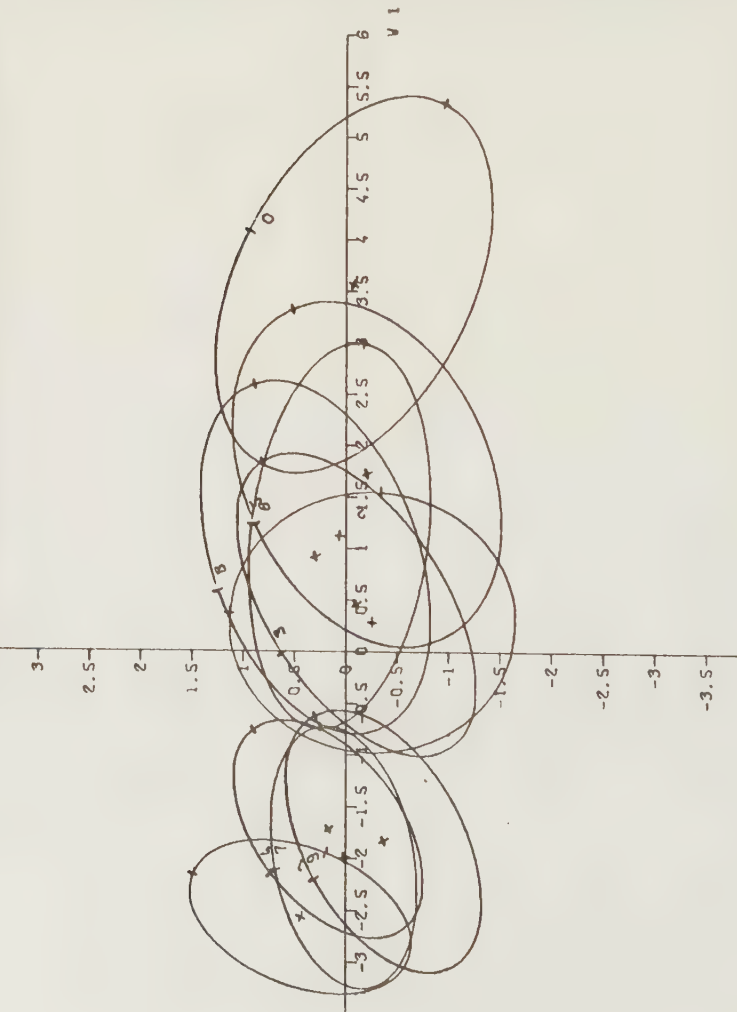
V 1W 3

KENNUNG 0

MITTELPUNKT, 3.57 -1.27

DK =2.00 A =2.08 B =0.90

3.5 v_{10} REQUIVALENTE ELLIPSEN $DK=KONST.$



DS-PROBE

HÖHNTR

V 1V10 KENNUNG 1

MITTELPUNKT, -2.57 0.43

DK =2.00 A =1.13 B =0.66

V 1V10 KENNUNG 2

MITTELPUNKT, 0.29 -0.26

DK =2.00 A =1.39 B =1.20

V 1V10 KENNUNG 3

MITTELPUNKT, 0.46 -0.10

DK =2.00 A =1.65 B =0.87

V 1V10 KENNUNG 4

MITTELPUNKT, -1.71 0.17

DK =2.00 A =1.20 B =0.72

V 1V10 KENNUNG 5

MITTELPUNKT, 1.73 -0.20

DK =2.00 A =1.75 B =1.20

V 1V10 KENNUNG 6

MITTELPUNKT, 1.13 0.06

DK =2.00 A =1.89 B =0.86

V 1V10 KENNUNG 7

MITTELPUNKT, -1.99 0.02

DK =2.00 A =1.26 B =0.69

V 1V10 KENNUNG 8

MITTELPUNKT, 0.92 0.30

DK =2.00 A =1.77 B =1.01

V 1V10 KENNUNG 9

MITTELPUNKT, -1.83 -0.36

DK =2.00 A =1.37 B =0.78

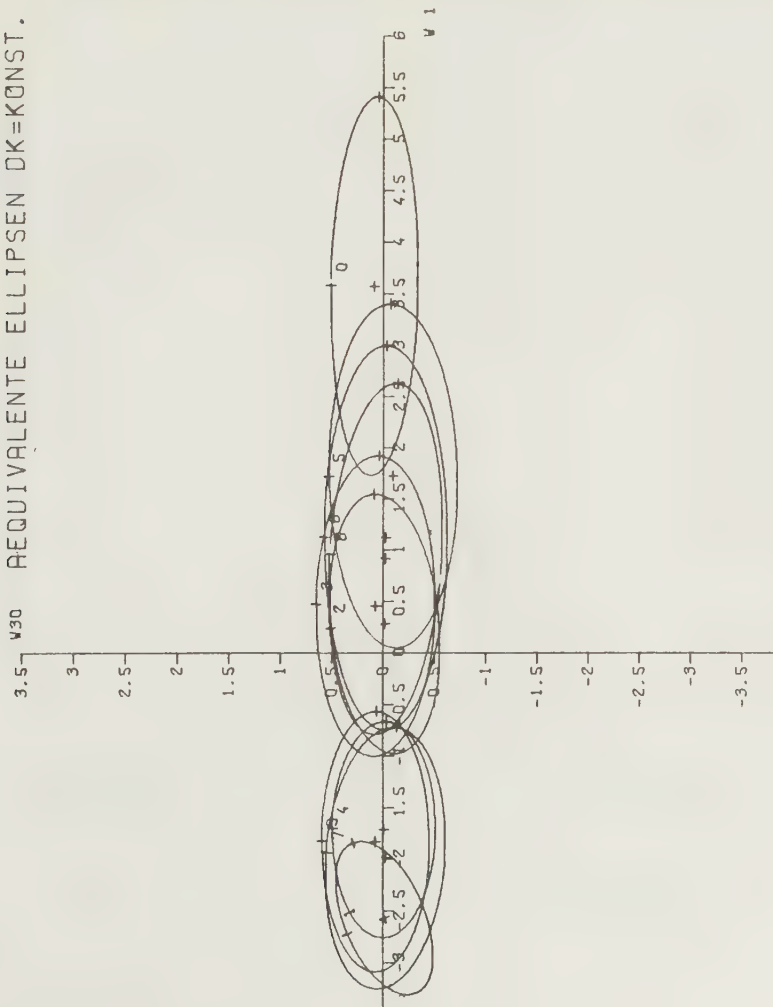
V 1V10 KENNUNG 0

MITTELPUNKT, 3.57 -0.07

DK =2.00 A =1.98 B =1.14

Bild 47. Äquivalente Ellipsen der 10 Klassen der Handblock-schriftstichprobe in der Ebene $w_1 w_{10}$.

W30 ÄQUIVALENTE ELLIPSEN DK=KONST.



OS-PROBE H6QHTA

V 1W30	KENNUNG 1
MITTELPUNKT, -2.57	-0.02
DK =2.00	A =0.78 B =0.40
V 1W30	KENNUNG 2
MITTELPUNKT, 0.29	-0.02
DK =2.00	A =1.26 B =0.53
V 1W30	KENNUNG 3
MITTELPUNKT, 0.46	0.07
DK =2.00	A =1.46 B =0.58
V 1W30	KENNUNG 4
MITTELPUNKT, -1.71	-0.01
DK =2.00	A =1.05 B =0.50
V 1W30	KENNUNG 5
MITTELPUNKT, 1.73	-0.10
DK =2.00	A =1.67 B =0.63
V 1W30	KENNUNG 6
MITTELPUNKT, 1.13	-0.03
DK =2.00	A =1.87 B =0.60
V 1W30	KENNUNG 7
MITTELPUNKT, -1.99	-0.03
DK =2.00	A =1.27 B =0.58
V 1W30	KENNUNG 8
MITTELPUNKT, 0.92	-0.02
DK =2.00	A =1.71 B =0.54
V 1W30	KENNUNG 9
MITTELPUNKT, -1.83	0.07
DK =2.00	A =1.26 B =0.52
V 1W30	KENNUNG 0
MITTELPUNKT, 3.57	0.08
DK =2.00	A =1.84 B =0.42

Bild 48. Äquivalente Ellipsen der 10 Klassen der Handblock-schriftstichprobe in der Ebene w_{130} .

WERTEVERTEILUNG DER ABSTANDSMASSE (REQ)

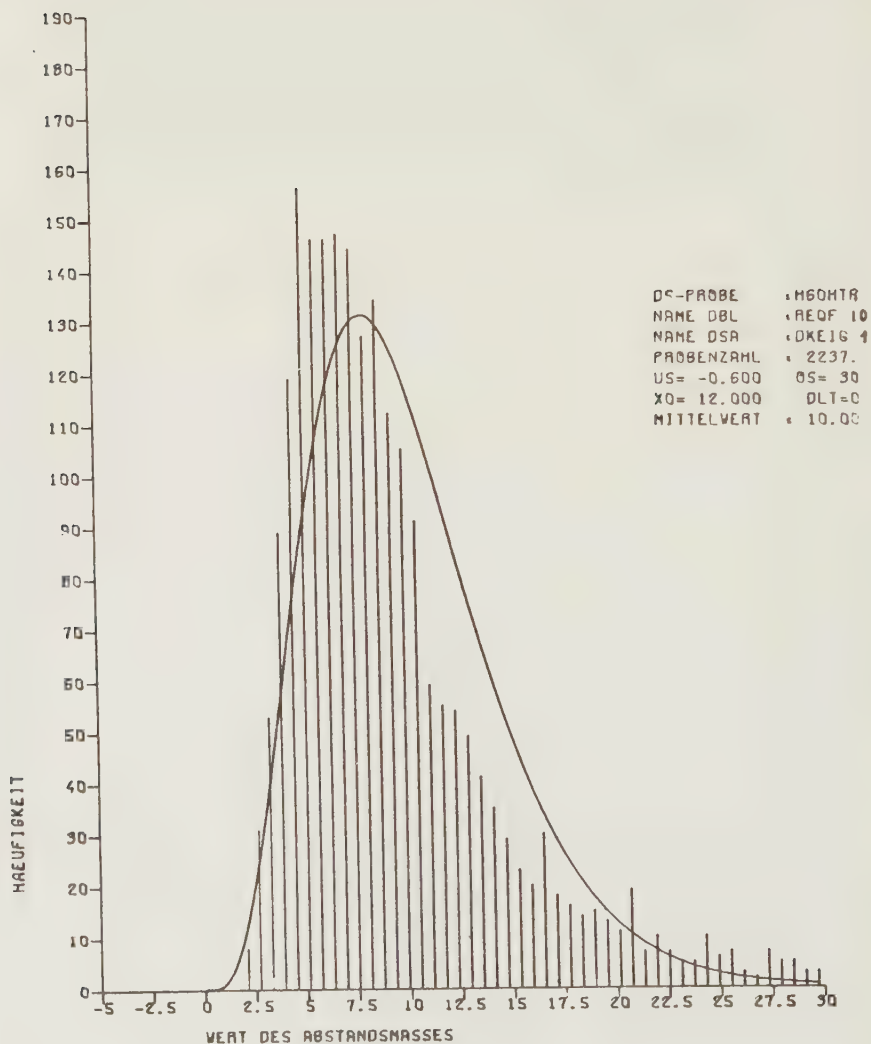


Bild 49. Werteverteilung von $Q_{k=1} \cdot \mathbf{w}$ in 10 Dimensionen.
 Stichprobe H0.

WERTEVERTEILUNG DER ABSTANDSMASSE (REQ)

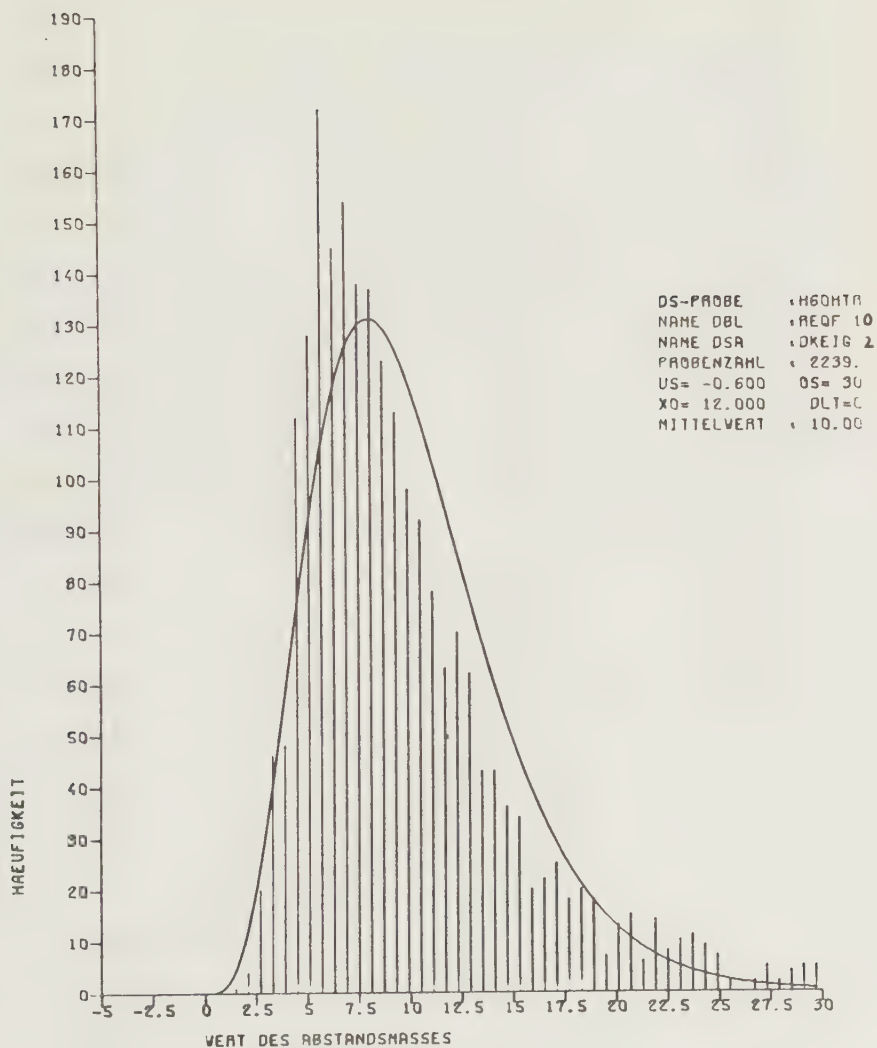


Bild 50. Werteverteilung von $Q_{k=2} \cdot \mathbf{w}$ in 10 Dimensionen.
Stichprobe H0.

WERTEVERTEILUNG DER ABSTANDSMASSE (REQ)

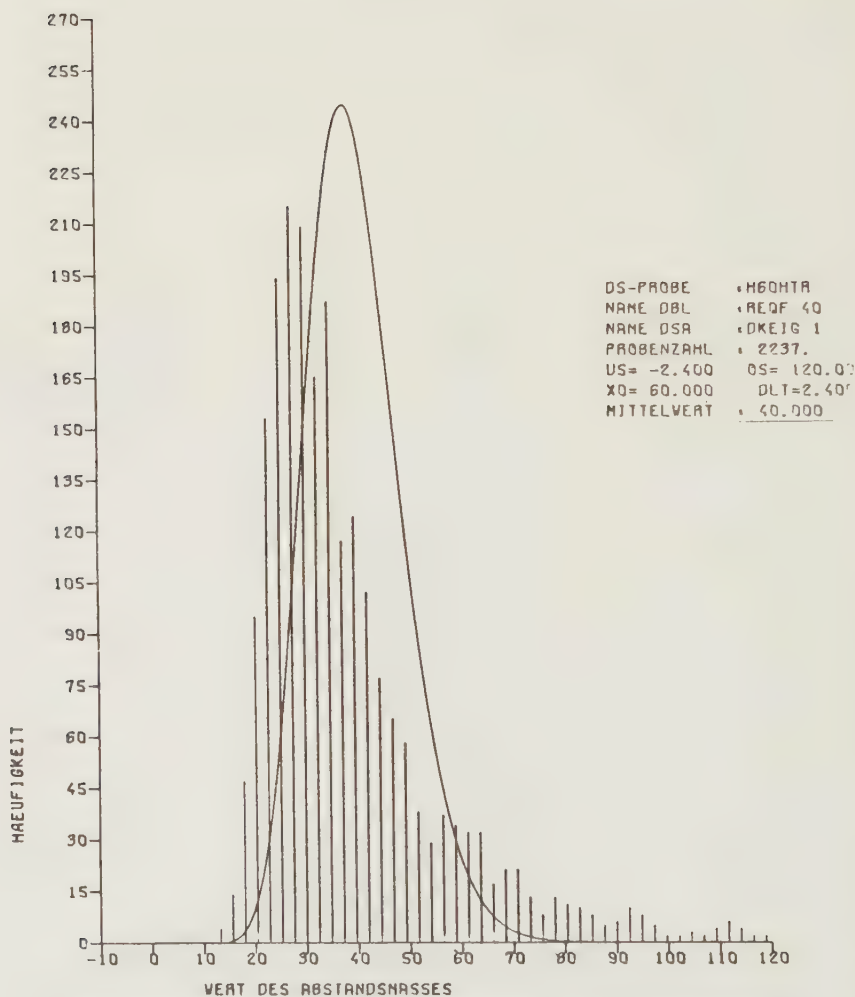


Bild 51. Werteverteilung von $Q_{k=1} \cdot W$ in 40 Dimensionen.
 Stichprobe H0.

WERTEVERTEILUNG DER ABSTANDSMASSE (REQ)

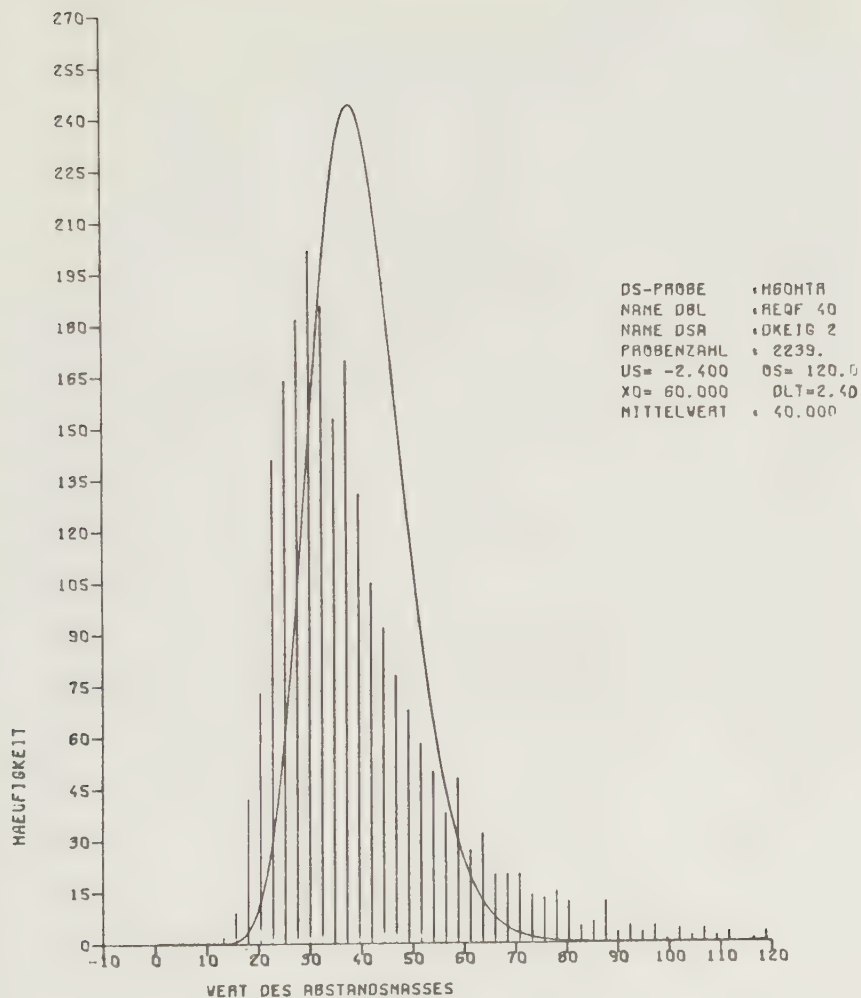
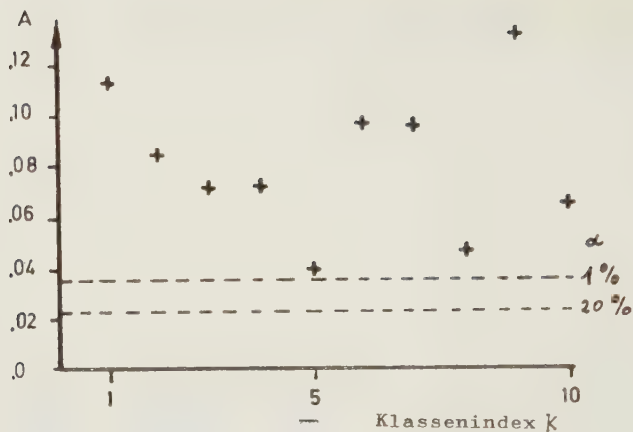
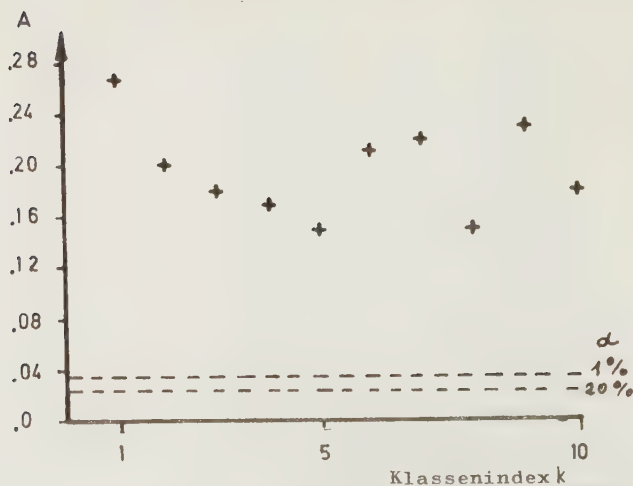


Bild 52. Werteverteilung von $Q_{k=2} \cdot \mathbf{w}$ in 40 Dimensionen.
 Stichprobe H0.



Dimension des Merkmalsvektors $\mathbf{W}$: $M = 10$.

2 200 Repräsentanten pro Klasse.



Dimension des Merkmalsvektors $\mathbf{W}$: $M = 40$.

2 200 Repräsentanten pro Klasse.

Bild 53. Ergebnisse des Kolmogoroff-Smirnow-Tests für die Verteilung des Abstandsmaßes Q_k der Handblock-schrift-Stichprobe H0. Modellkurve: χ^2 -Verteilung.

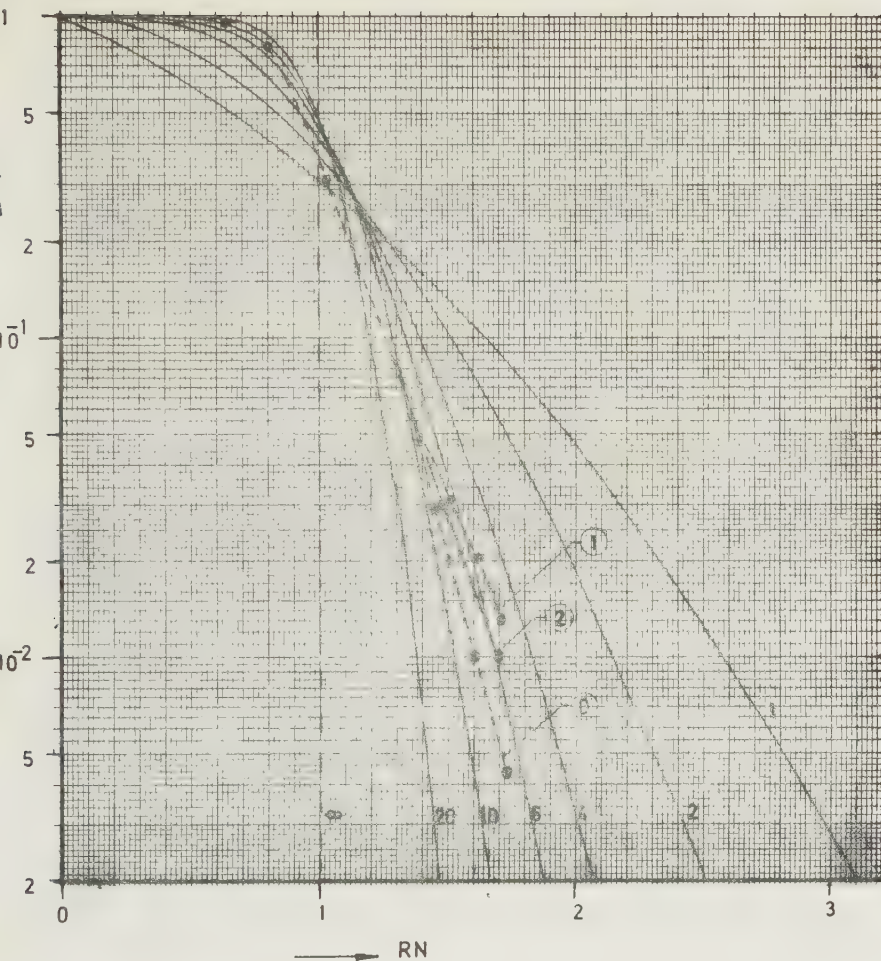


Bild 54. Anteil a der Zeichen außerhalb des Hyperellipsoids $Q_k = RN^2 \cdot M$ beim Modell der M-dimensionalen Normalverteilung (ausgezogene Kurven, gezeichnet für $M = 1, 2, 4, 6, 10, 20$) und bei den ausgemessenen Verteilungen der Merkmale (40-dimensionaler Raum) von Zeichen mit den Kennungen 0, 1, 2 (unterbrochene Kurven).

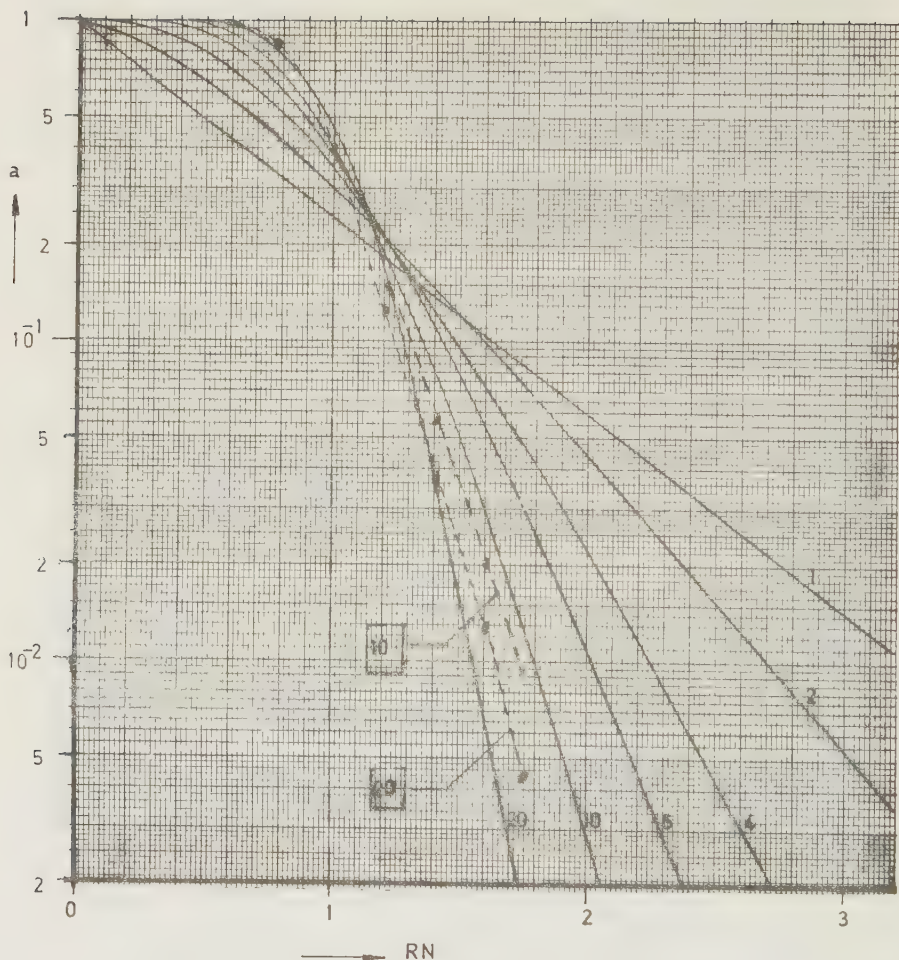
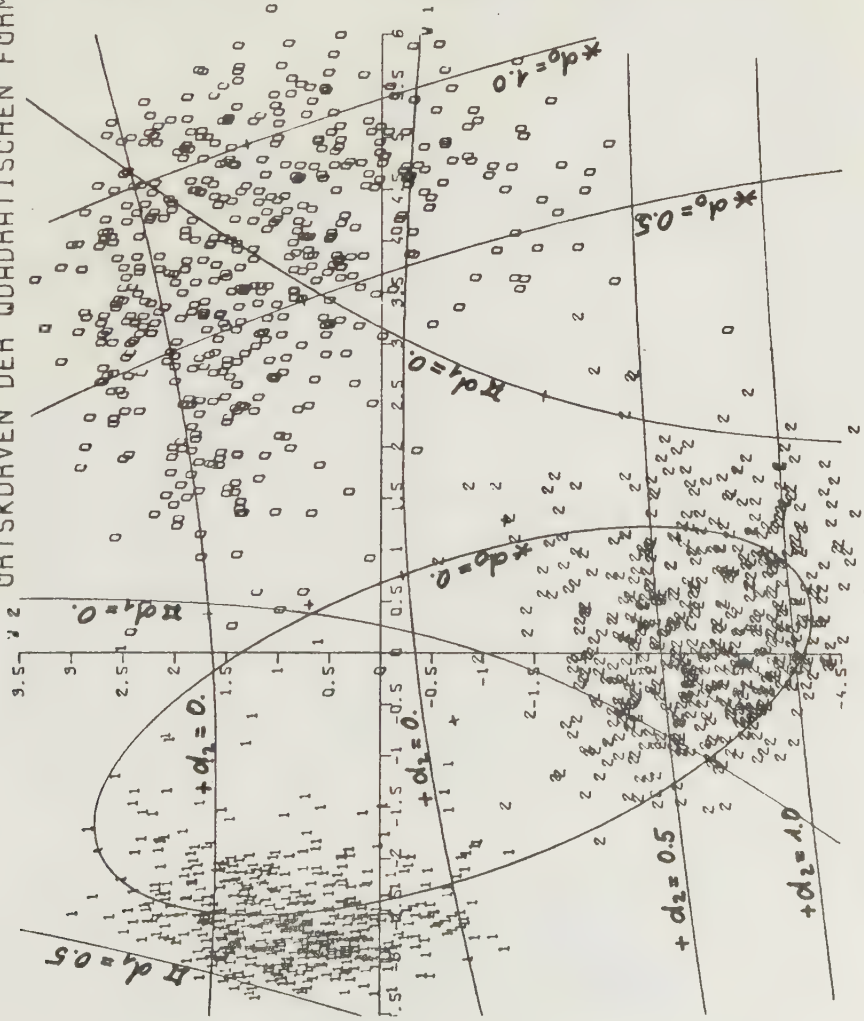


Bild 55. Anteil a der Zeichen außerhalb des Hyperellipsoids $Q_k = RN^2 \cdot M$ beim Modell der M -dimensionalen Laplaceverteilung (ausgezogene Kurven, gezeichnet für $M = 1, 2, 4, 6, 10, 20$) und bei den ausgemessenen Verteilungen der Merkmale (40 , und 10 dimensionaler Raum) von Zeichen der Kennung "0" (unterbrochene Kurven).

ORTSKURVEN DER QUADRATISCHEN FORM



DS-PROBE H60HTA

→→ MESSWERTE IN EBENE V1, V2.

KENNUNG 1 PROBENZAHL, 500.

KENNUNG 2 PROBENZAHL, 500.

KENNUNG 0 PROBENZAHL, 500.

V1 V2 KENNUNG 1

MITTELPUNKT, 1.28 -1.22

DK = 0.00 A = 1.27 B = 3.28

DK = 0.50 A = 4.99 B = 12.85

DK = 1.00 A = 6.94 B = 17.67

V1 V2 KENNUNG 2

MITTELPUNKT, 0.46 0.69

DK = 0.00 A = 0.98 B = 4.45

DK = 0.50 A = 3.38 B = 15.28

DK = 1.00 A = 4.68 B = 21.15

V1 V2 KENNUNG 0

MITTELPUNKT, -0.65 -0.73

DK = 0.00 A = 1.49 B = 3.88

DK = 0.50 A = 4.33 B = 10.65

DK = 1.00 A = 5.94 B = 14.61

Bild 56.

Ortskurven des Quadrat-
mittelklassifikators
mit konstanter Schät-

zung d_k der Klassen-
zugehörigkeit.

Ebene $w_1 w_2$.

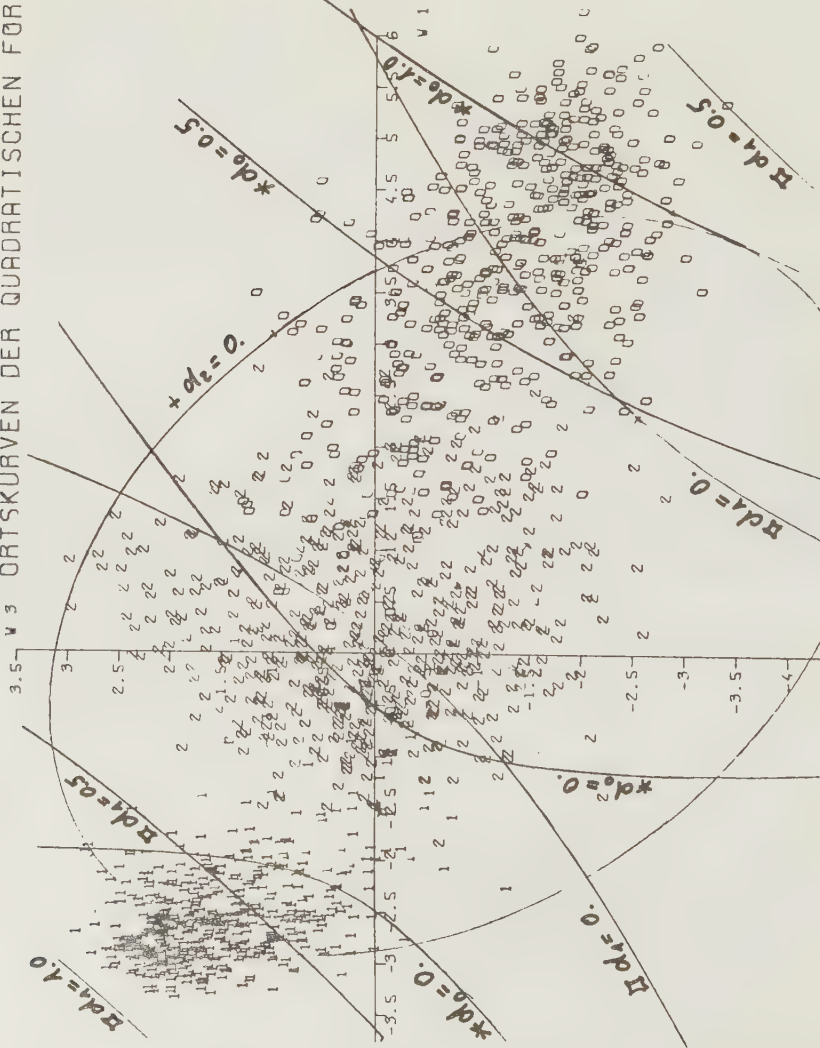
10 Klassen gelernt.

* Kennung "0"

□ Kennung "1"

+ Kennung "2"

V 3 ÖRTSKURVEN DER QUADRATISCHEN FORM



OS-PROBE

HÖCHSTA

+++ MESSWEITE IN EBENE V1. V3.

KENNUNG 1 PROBENZAHL: 500.

KENNUNG 2 PROBENZAHL: 500.

KENNUNG 0 PROBENZAHL: 500.

V 1 V 3 KENNUNG 1

MITTELPUNKT: 1.07 -1.49

DK = 0.07 $\bar{x}$ = 1.63 $\bar{y}$ = 3.43

DK = 0.50 $\bar{x}$ = 4.43 $\bar{y}$ = 9.32

DK = 1.00 $\bar{x}$ = 6.04 $\bar{y}$ = 12.75

V 1 V 3 KENNUNG 2

MITTELPUNKT: 6.57 -0.32

DK = 0.00 $\bar{x}$ = 3.02 $\bar{y}$ = 4.43

V 1 V 3 KENNUNG 0

MITTELPUNKT: -1.53 -0.06

DK = 0.00 $\bar{x}$ = 0.34 $\bar{y}$ = 1.50

DK = 0.50 $\bar{x}$ = 4.61 $\bar{y}$ = 6.23

DK = 1.00 $\bar{x}$ = 6.49 $\bar{y}$ = 11.55

Bild 57.

Ortskurven des Quadratt-
mittelklassifikators
mit konstanter Schät-
zung d_k der Klassen-
zugehörigkeit.

Ebene w_{13} .

10 Klassen gelernt

* Kennung "0"

⊠ Kennung "1"

+ Kennung "2"

ORTSKURVEN DER QUADRATISCHEN FORM

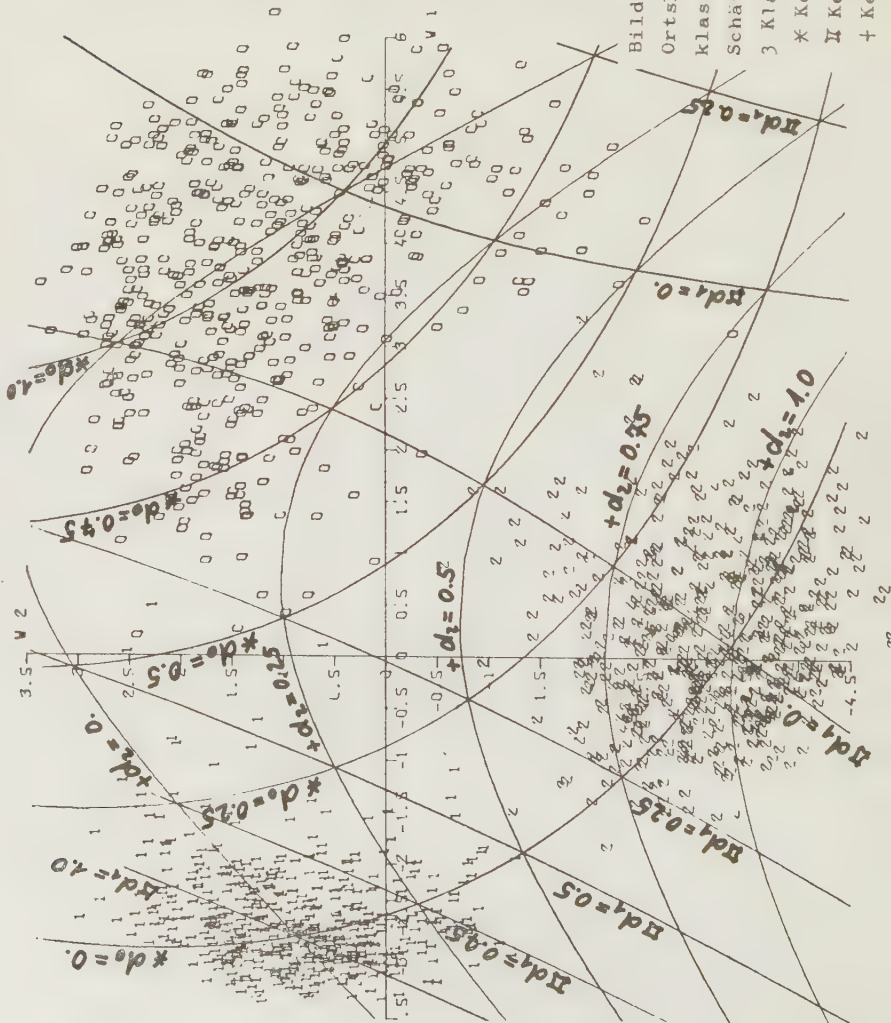


Bild 59.

Ortskurven des Quadratmittel-
klassifikators mit konstanter

Schätzung d_k . Ebene $w_1 w_2$.

3 Klassen gelernt

* Kennung "0"

II Kennung "1"

+ Kennung "2"

OS-PROBE HCHOHTA
+++ MESSVERTE IN EBENE v_1, v_2
KENNUNG 1 PROBENZAHL, SOO.
KENNUNG 2 PROBENZAHL, SCO.
KENNUNG 3 PROBENZAHL, SCC.

$v_1 v_2$ KENNUNG 1
MITTELPUNKT, 3.47 0.49

DK = 0.00 A = 1.00 B = 3.25
DK = 0.25 A = 3.07 B = 9.91
DK = 0.50 A = 4.22 B = 13.64
DK = 0.75 A = 5.12 B = 16.57
DK = 1.00 A = 5.88 B = 19.00

$v_1 v_2$ KENNUNG 2

MITTELPUNKT, 2.10 0.84
DK = 0.00 A = 2.22 B = 3.52
DK = 0.25 A = 3.37 B = 5.93
DK = 0.50 A = 5.15 B = 7.70
DK = 0.75 A = 6.11 B = 7.73
DK = 1.00 A = 6.94 B = 10.37

$v_1 v_2$ KENNUNG 3

MITTELPUNKT, 263.63 143.54
DK = 0.00 A = 42.46 B = 302.17
DK = 0.25 A = 42.15 B = 300.95
DK = 0.50 A = 41.91 B = 299.41
DK = 0.75 A = 41.72 B = 297.97
DK = 1.00 A = 41.50 B = 296.50

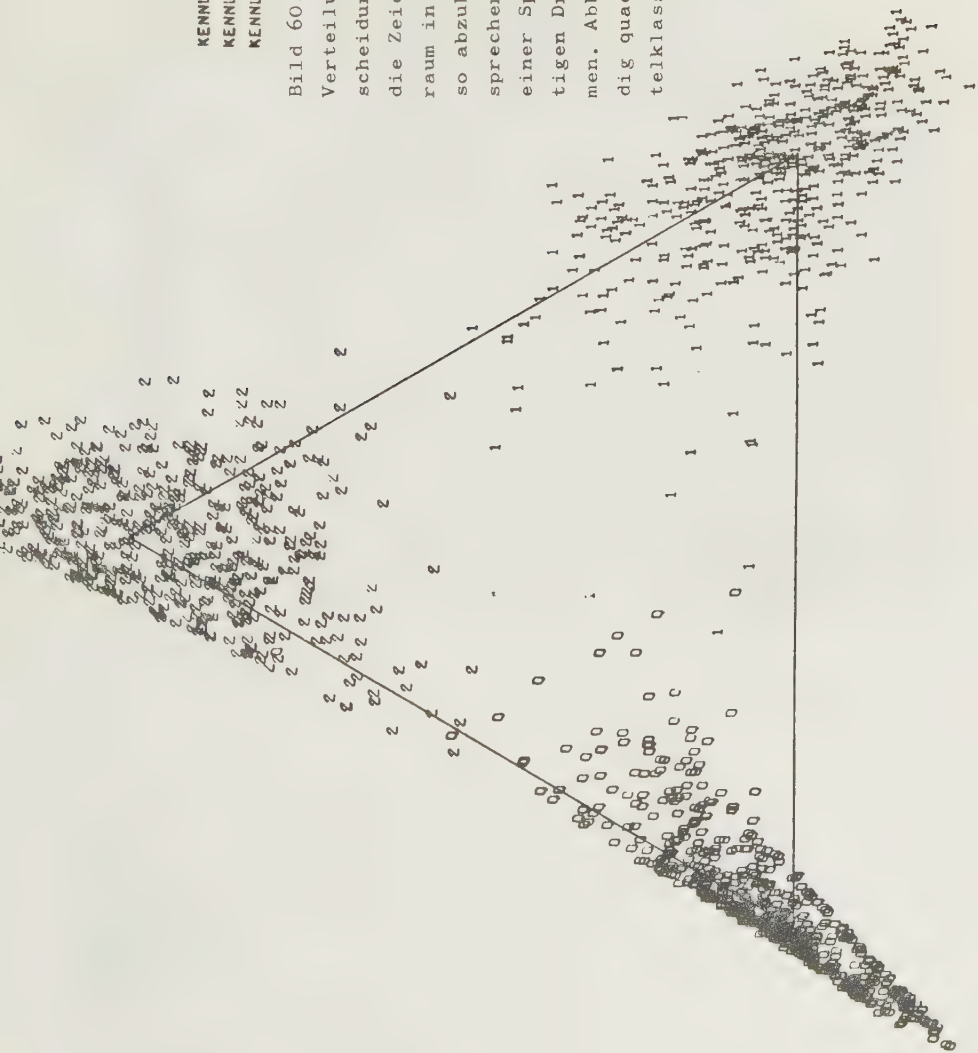
KENNUNG 1 PROBENZAHL, 500.

KENNUNG 2 PROBENZAHL, 500.

KENNUNG 0 PROBENZAHL, 500.

Bild 60.

Verteilung der Zeichen im Entscheidungsräum. Ziel ist es, die Zeichen von dem Merkmalsraum in den Entscheidungsraum so abzubilden, daß sie, entsprechend ihrer Kennung, an einer Spitze des gleichseitigen Dreiecks zu liegen kommen. Abbildung durch vollständige quadratischen Quadratmittelklassifikator.



KENNUNG 1 PROBENZAHL: 500.
 KENNUNG 2 PROBENZAHL: 500.
 KENNUNG C PROBENZAHL: 500.

Bild 61.

Verteilung der Zeichen im
 Entscheidungsraum. Die Ab-
 bildung vom Merkmalsraum in
 den Entscheidungsraum er-
 folgte mit dem in 2 Merk-
 malen vollständig kubischen
 Quadratmittelklassifikator.



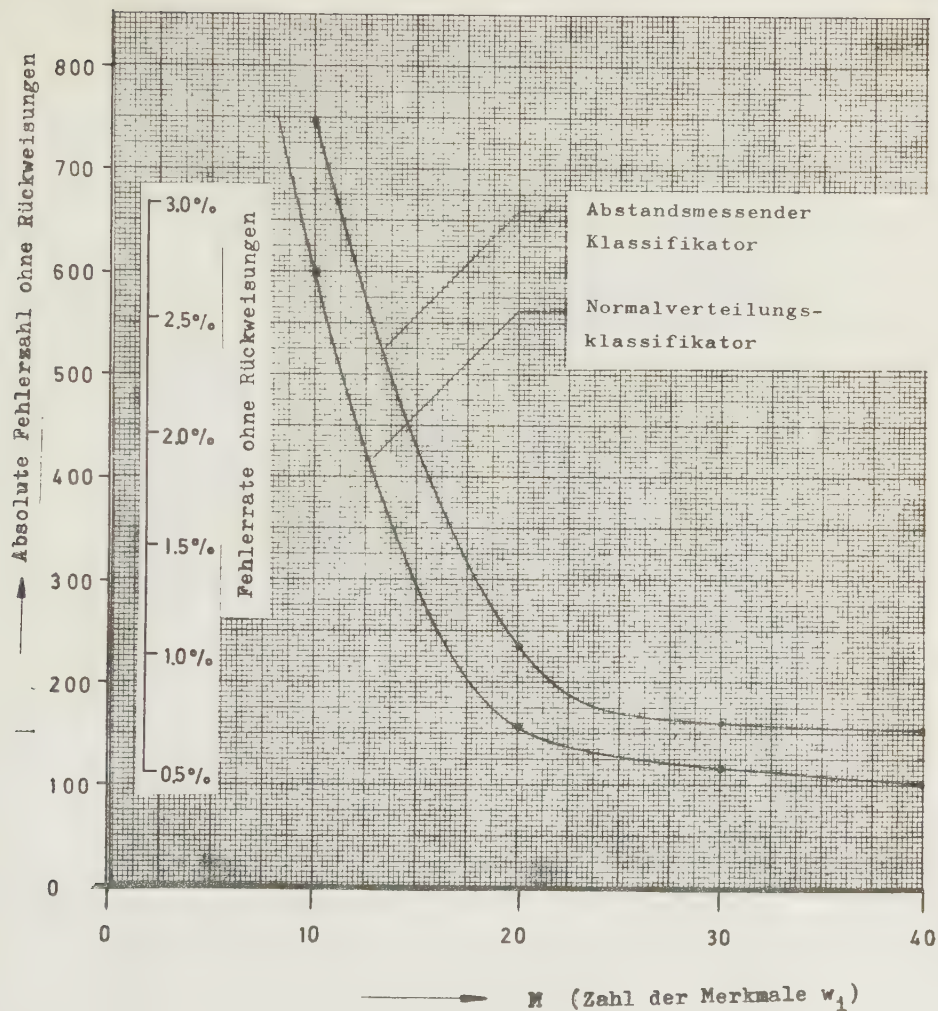


Bild 62. Absolute Fehlerzahl von abstandsmessendem Klassifikator und Normalverteilungsklassifikator für die Handblockschrift-Stichprobe H0 (22 000 Zeichen).

KENNUNG	ZAHL M DER MERKMALE W			
	40	30	20	10
1	1.5E-43	6.0E-31	2.0E-17	5.0E-07
2	1.4E-33	1.5E-22	1.0E-12	1.3E-04
3	6.0E-29	6.0E-19	2.4E-10	2.0E-03
4	1.5E-32	1.5E-21	4.6E-12	1.0E-04
5	3.0E-27	1.4E-17	1.1E-09	1.0E-02
6	3.0E-32	2.6E-21	7.0E-12	3.0E-04
7	3.0E-35	4.5E-24	6.0E-14	1.0E-05
8	2.5E-26	4.8E-17	4.0E-09	7.0E-03
9	8.0E-34	3.4E-21	2.0E-12	1.2E-04
0	6.0E-30	3.3E-20	1.0E-16	9.0E-03

Bild 63. Die Determinanten der klassenspezifischen Kovarianzmatrizen $K_k = E \{ (\mathbf{w} - \bar{\mathbf{w}}_k) (\mathbf{w} - \bar{\mathbf{w}}_k)^T \}$ für die Handblockschrift-Stichprobe in Abhängigkeit von der Dimension M des Merkmalsvektors $\mathbf{w}$.

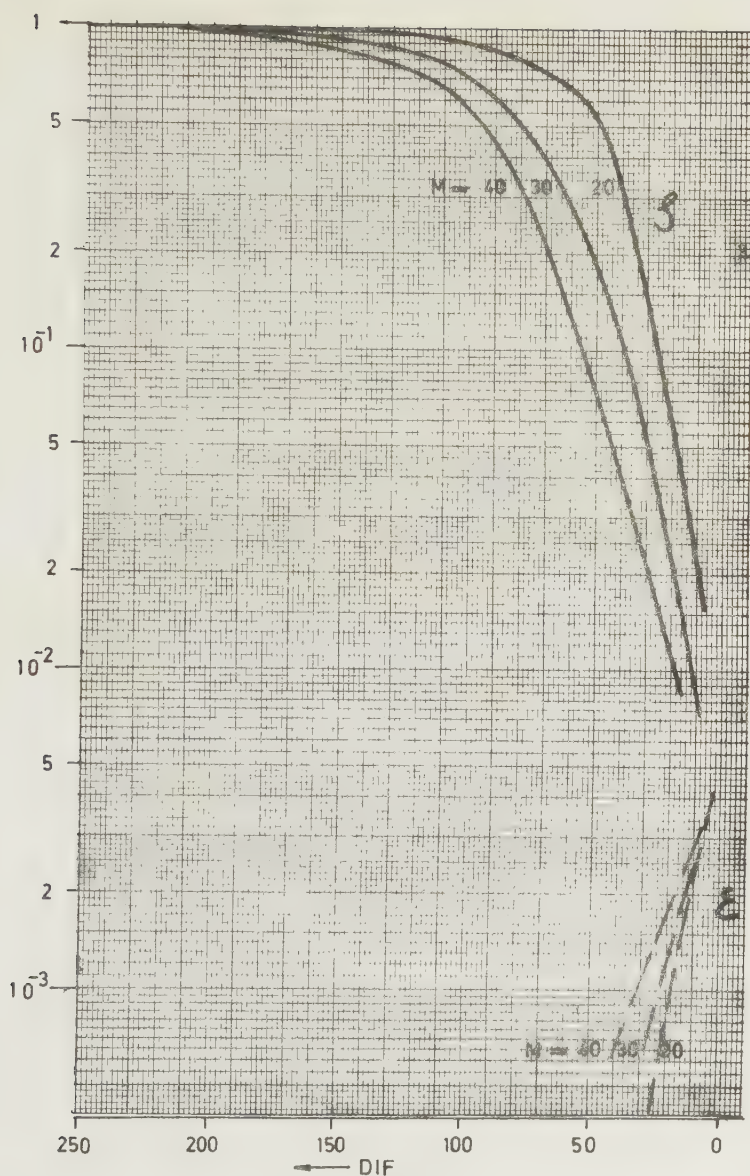


Bild 64. Zurückweisungsrate β und Fehlerrate ϵ für den Normalverteilungsklassifikator mit M Merkmalen w_i , abgetragen über der Differenz DIF zwischen bester und zweitbesten Schätzung.

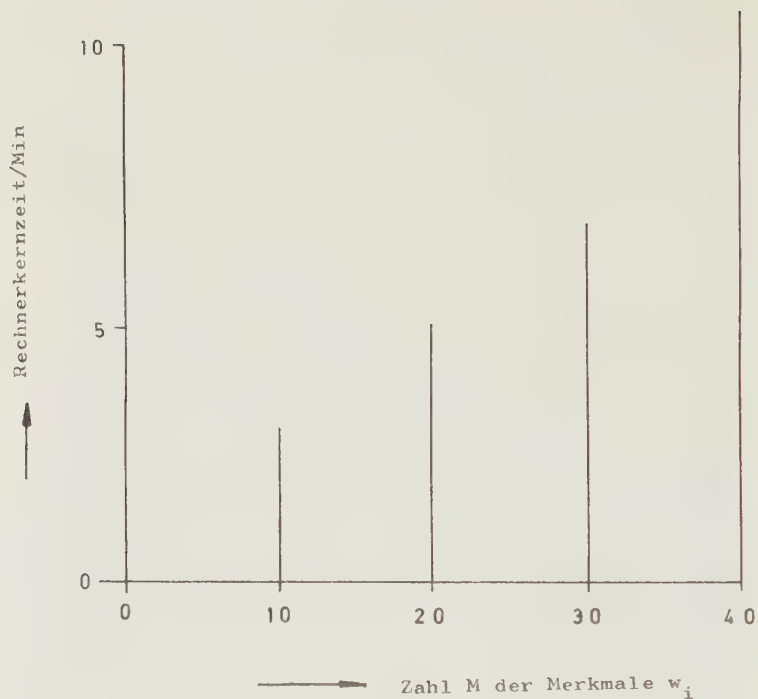


Bild 65. Rechnerkernzeit, die zur Berechnung des Klassifikators notwendig ist, und ein Maß für den Adaptionsaufwand darstellt, abgetragen über der Zahl M der verwendeten Merkmale w_i . Die Werte sind gültig für abstandsmessenden und Normalverteilungsklassifikator. Verwendete Stichprobe: H0.

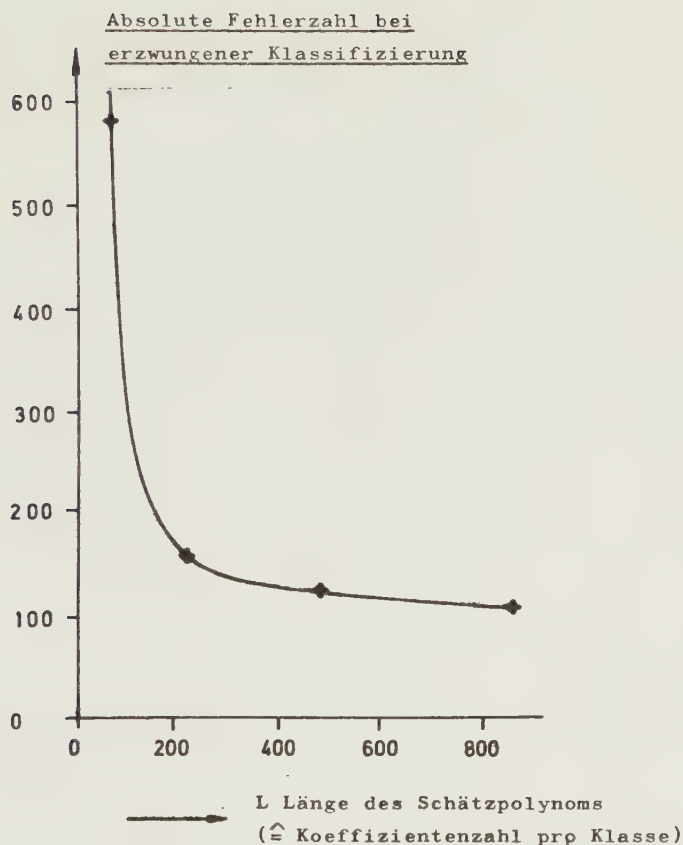


Bild 66. Absolute Fehlerzahl bei erzwungener Klassifizierung, abgetragen über der Länge L des Schätzpolynoms. Die Kurve wurde für Normalverteilungsklassifikator (Stichprobe H0) gemessen.

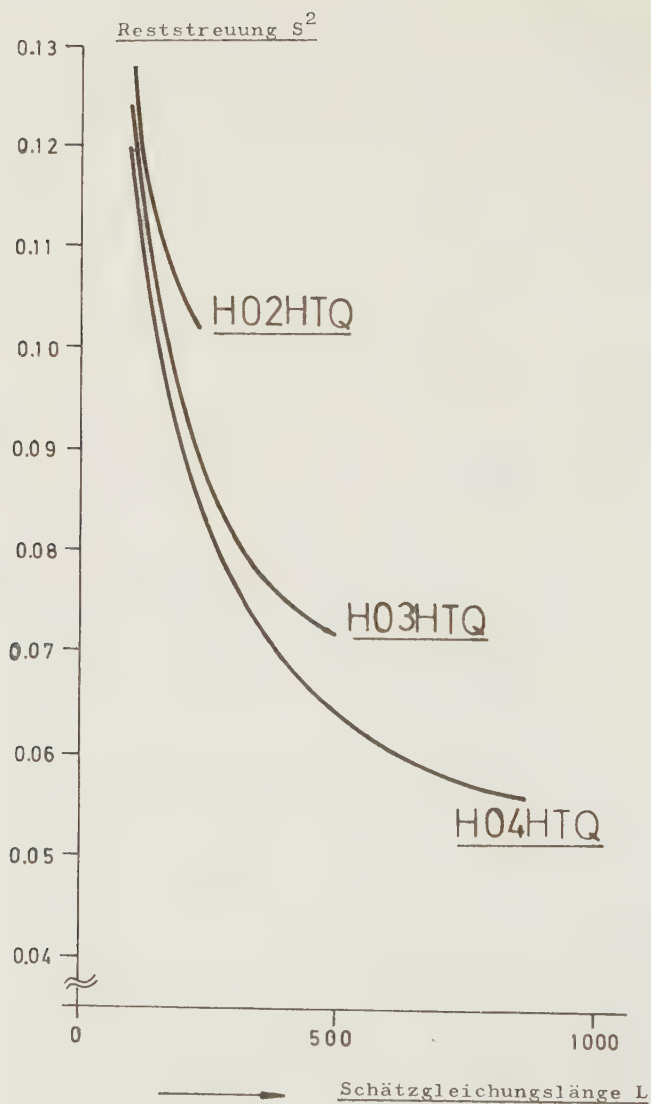


Bild 67. Reststreuungsdiagramme von 3 Quadratmittelklassifikatoren (Stichprobe H0).

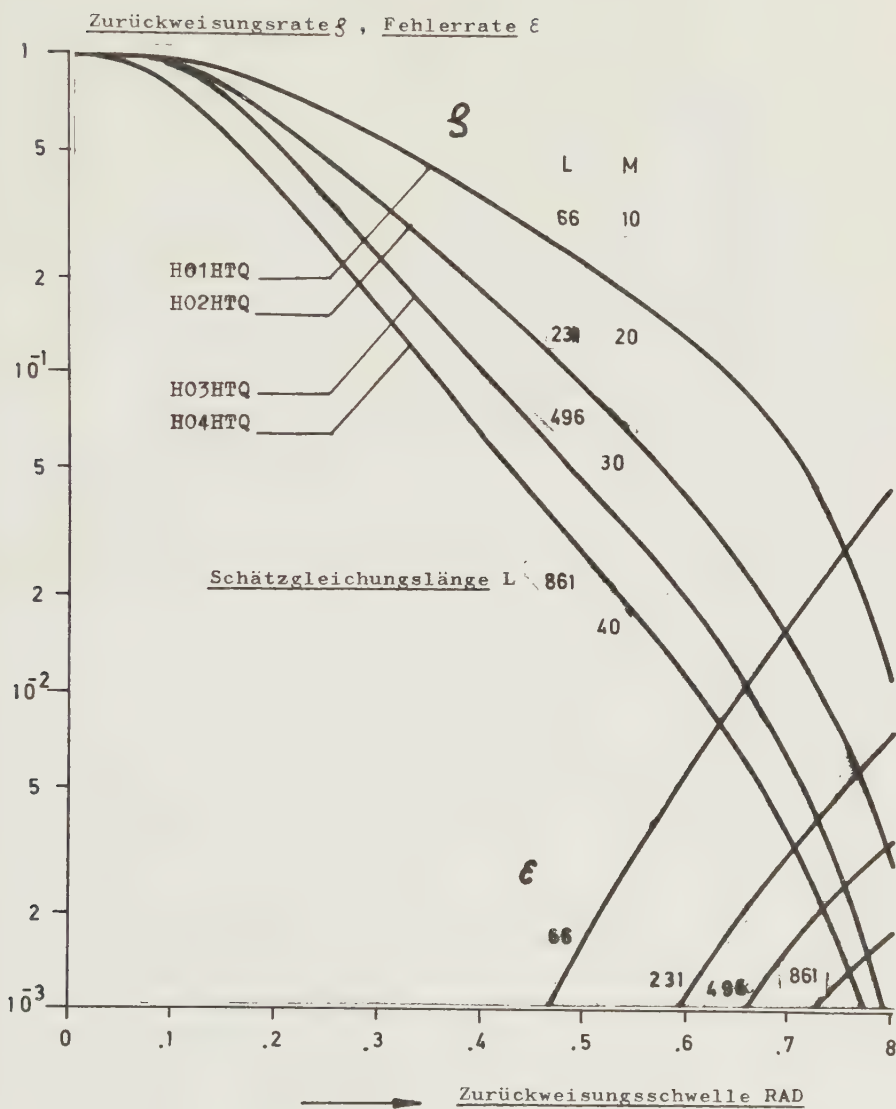


Bild 68. Rückweisungs- und Fehlerkurven des Quadratmittelklassifikators mit vollständig quadratischen Schätzpolynomen.

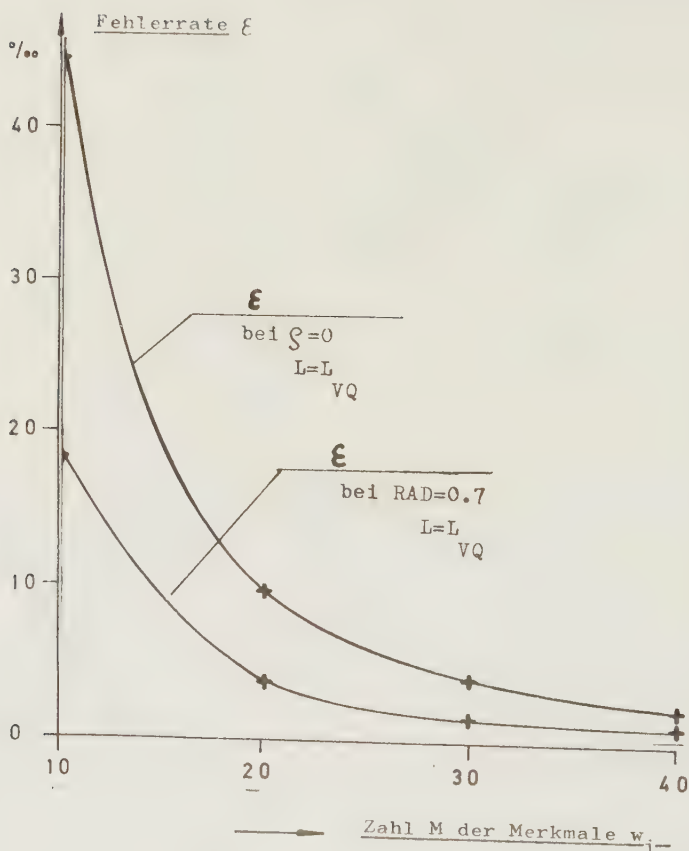


Bild 69. Erreichte Fehlerrate ohne Rückweisungen sowie Fehlerrate bei konstanter Rückweisungsschwelle $RAD = 0.7$, abgetragen über der Zahl der verwendeten Merkmale w_i . Das Schätzpolynom ist in allen Fällen vollständig quadratisch. Gültig für Quadratmittelklassifikatoren.

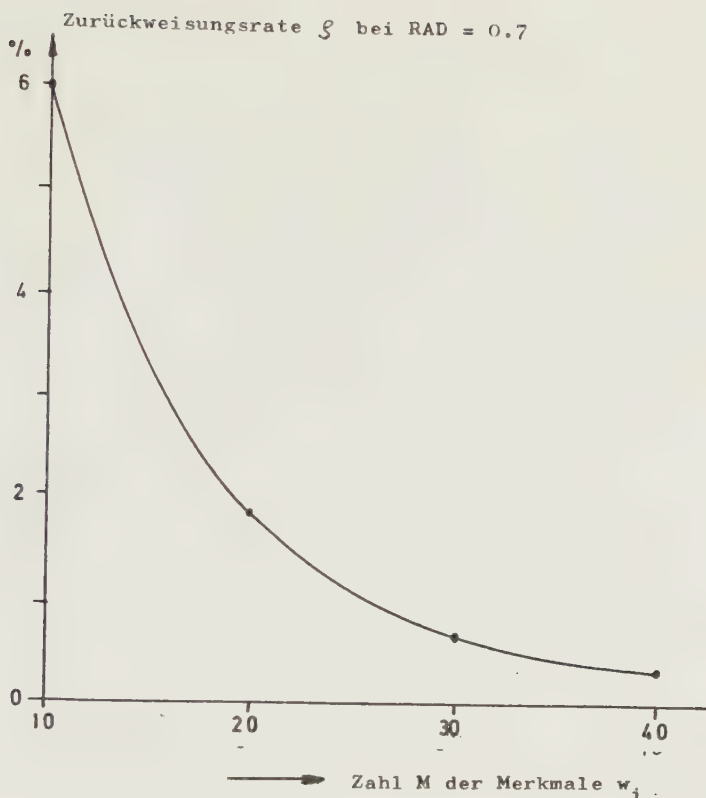


Bild 70. Erreichte Zurückweisungsrate ξ bei konstanter Zurückweisungsschwelle RAD = 0.7, abgetragen über der Zahl der verwendeten Merkmale w_i . Das Schätzpolynom ist in allen Fällen vollständig quadratisch.

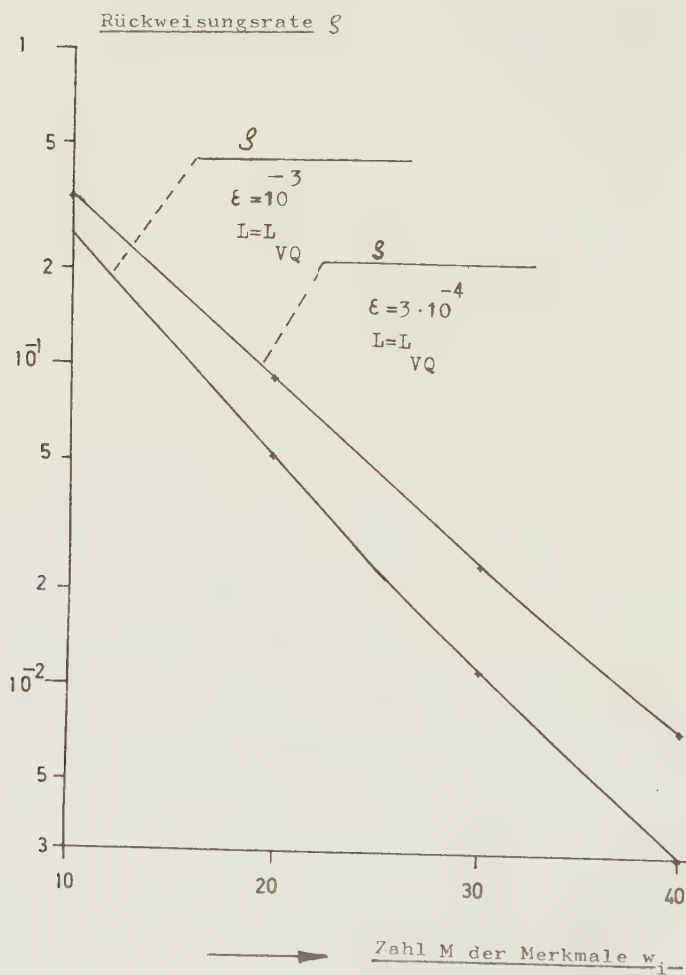


Bild 71. Rückweisungen, die notwendig sind, um verschiedene Fehlerraten zu erreichen, abgetragen über der Zahl der Merkmale w_i . Das Schätzpolynom ist in allen Fällen vollständig quadratisch. Kurven gültig für Quadratmittelklassifikatoren.

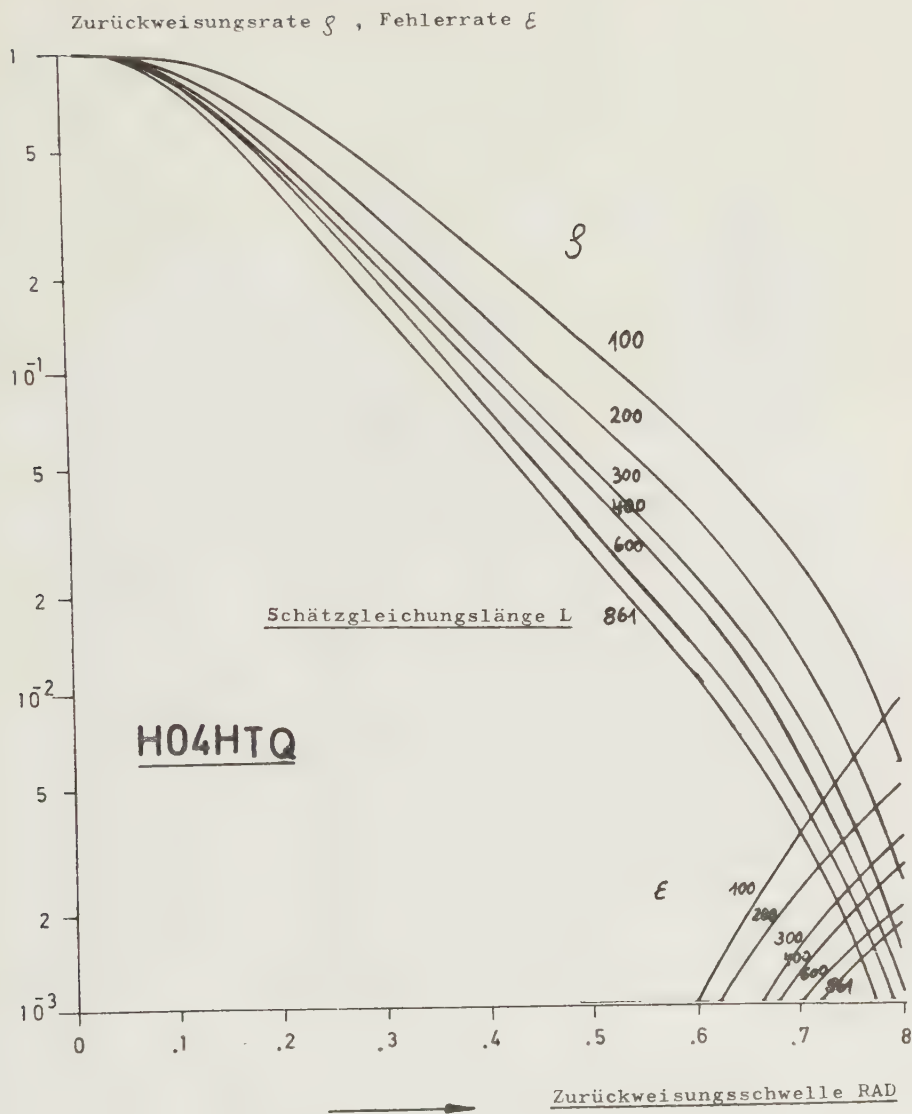


Bild 72. Zurückweisungs- und Fehlerkurven des Klassifikators HO4HTQ.

INDEX k

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	1	1	17												
2	1	40	22												
3	3	7	11	35											
4	4	13	46	23	54										
5	5	10	25	44	127	198									
6	6	39	14	73	19	785	41								
7	9	28	29	36	47	63	123	203							
8	8	21	18	55	112	829	69	169	153						
9	12	42	45	163	72	693	207	83	88	152					
10	16	32	38	60	110	332	86	287	232	289	101				
11	15	34	24	50	62	646	113	138	221	194	79	93			
12	27	31	49	274	156	518	116	197	192	315	111	552	202		
13	33	51	43	65	66	338	78	82	132	297	461	91	146	135	
14	84	57	58	172	37	582	204	68	482	342	175	187	186	147	99
15	26	160	67	53	106	105	822	386	288	337	181	459	523	264	313
16	30	365	61	201	1619	20	104	92	436	140	141	302	583	1267	354
17	52	145	85	179	190	97	228	164	255	94	114	193	284	115	471
18	136	401	139	324	1388	283	460	333	597	368	462	191	624	359	328
19	137	335	117	261	154	176	491	262	96	130	171	247	344	353	576
20	277	155	125	272	465	253	428	609	702	218	200	414	490	380	595

INDEX i

Bild 73. Die durch den Regressionsprozess ermittelte Rangreihenfolge (MS-Strategie) der Merkmale w_i (Spalte mit $k = 0$) und der Merkmalsverknüpfungen w_{ik} .

INDEX k

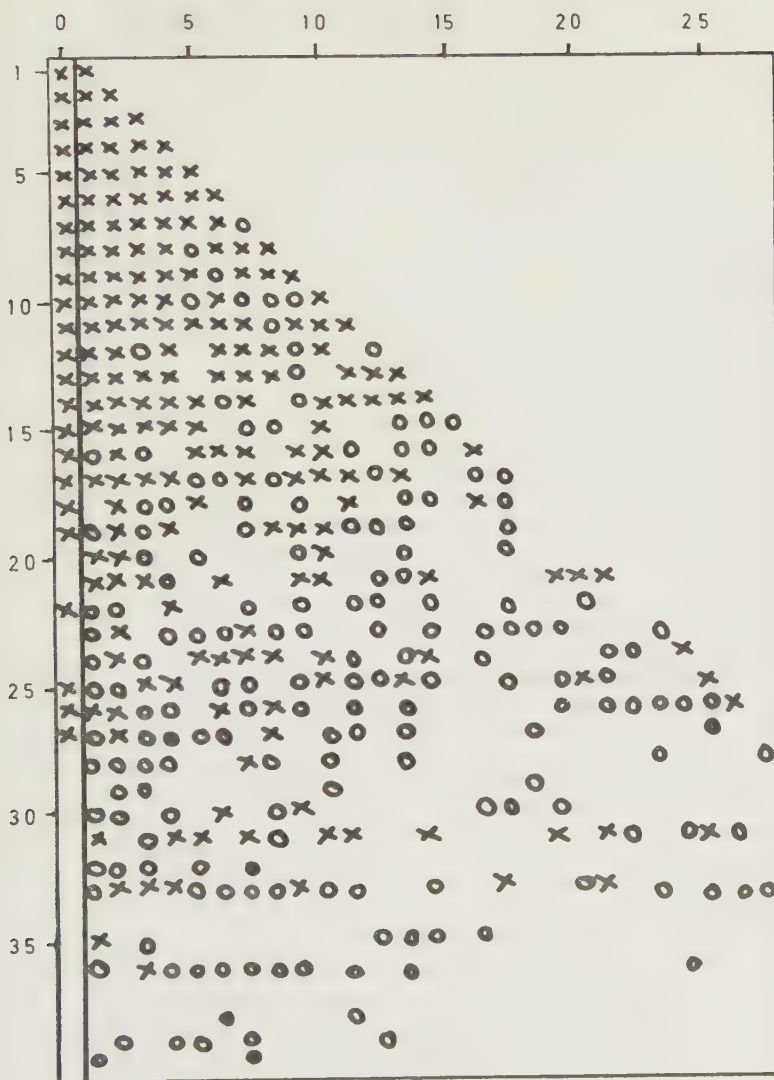


Bild 74. Graphische Darstellung der 200 (400) wichtigsten Merkmale bzw. Merkmalsverknüpfungen, ermittelt während des Regressionsprozesses (MS-Strategie).

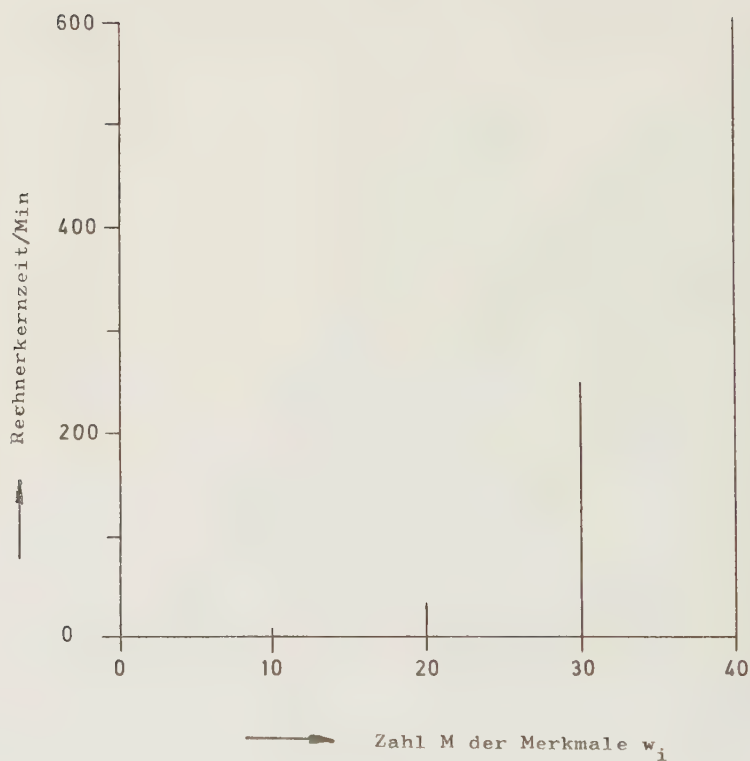


Bild 75. Rechnerkernzeit, die zur Berechnung des Quadratmittelklassifikators notwendig ist, und ein Maß für den Adaptionaufwand darstellt, abgetragen über der Zahl M der verwendeten Merkmale w_i . Die Merkmale w_i werden als Fortran REAL Variable definiert (39 BIT). Verwendete Stichprobe H0.

Absolute Fehlerzahl bei
erzwungener Klassifizierung

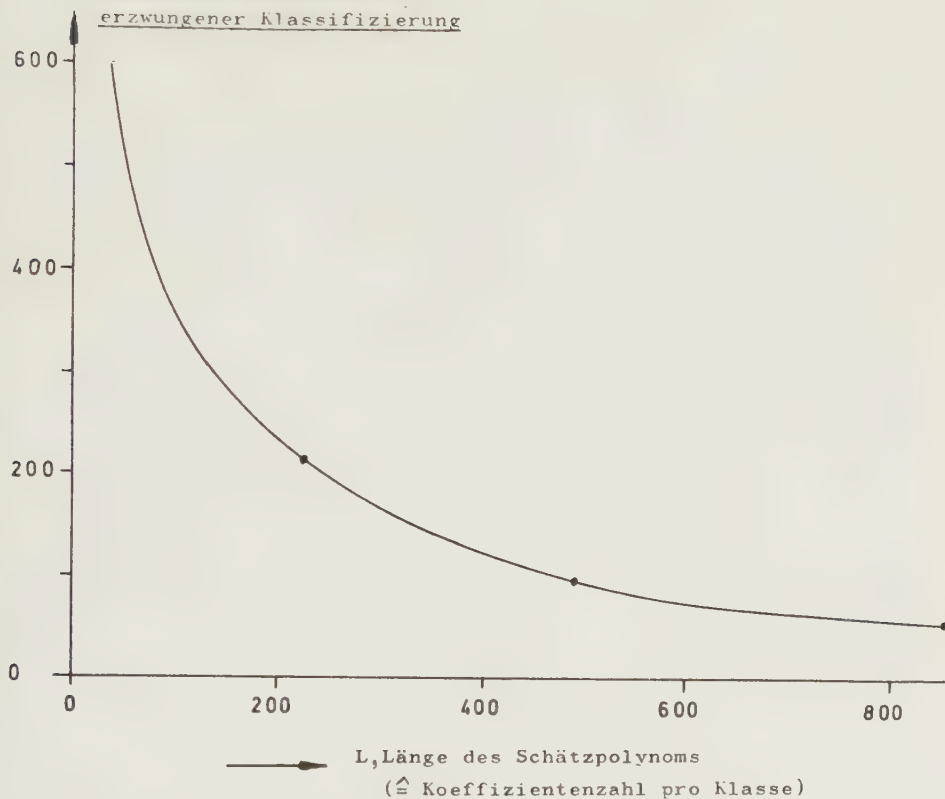


Bild 76. Absolute Fehlerzahl bei erzwungener Klassifizierung, abgetragen über der Länge L des Schätzpolynoms. Die Kurve wurde für den Quadratmittelklassifikator bei jeweils vollständig quadratischem Schätzpolynom gemessen. (Stichprobe H0).

Rückweisungsrate $\mathcal{S}$ bei $RAD \approx 0.7$

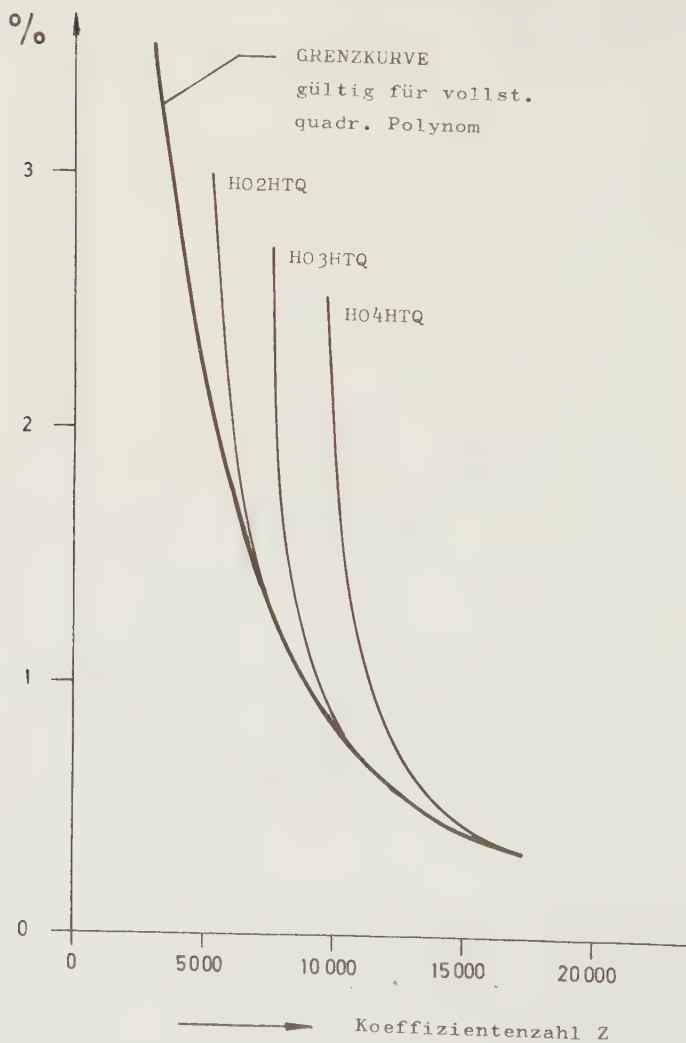


Bild 77. Erreichte Systemleistung (gemessen mit $\mathcal{S}_{RAD=0.7}$), abgetragen über dem Systemaufwand (gemessen mit der Koeffizientenzahl).

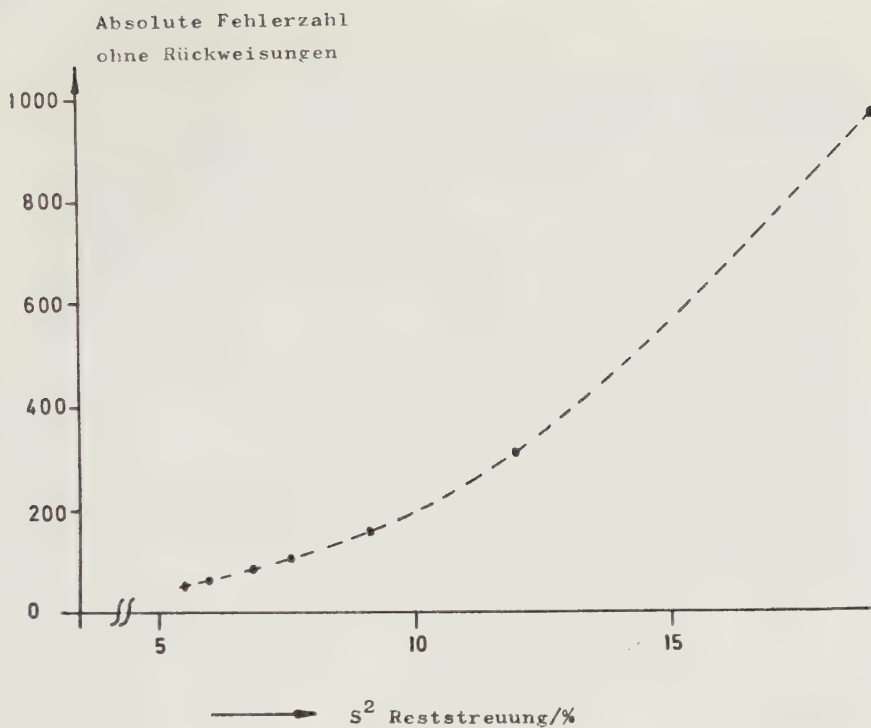


Bild 78. Absolute Fehlerzahl ohne Rückweisungen, abgetragen als Funktion der Reststreuung S^2 . Die Kurve wurde für den Quadratmittelklassifikator H04HTQ ($S^2 < 15\%$) und H01HTQ ($S^2 > 15\%$) gemessen.

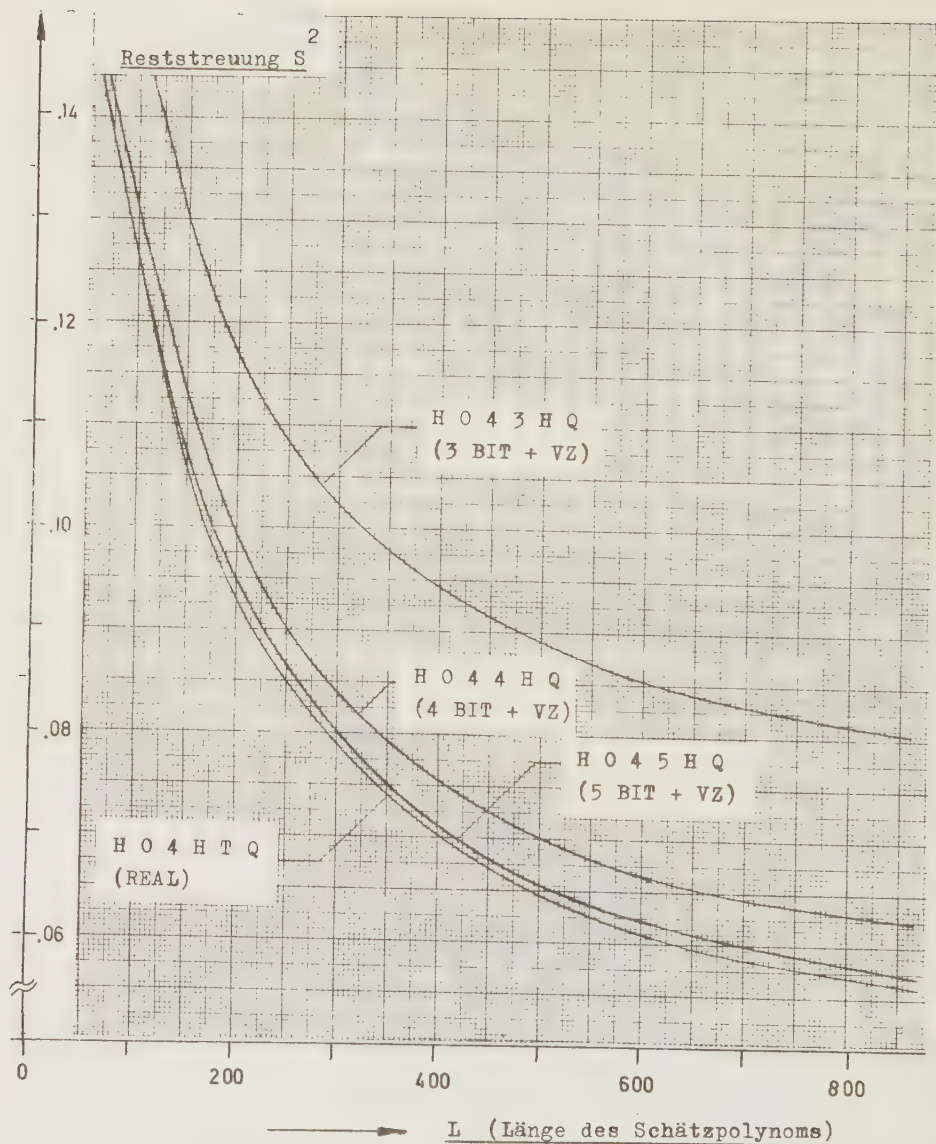


Bild 79. Erreichte Reststreuung S^2 der Klassifikatoren HO4HTQ (REAL-Darstellung der w_i), HO45HQ (w_i mit 5 BIT + Vorzeichen quantisiert), HO44HQ (w_i mit 4 BIT + VZ), HO43HQ (w_i mit 3 BIT + VZ), abgetragen über der Länge L des Schätzpolynoms.

Zurückweisungsrate ξ , Fehlerrate ϵ

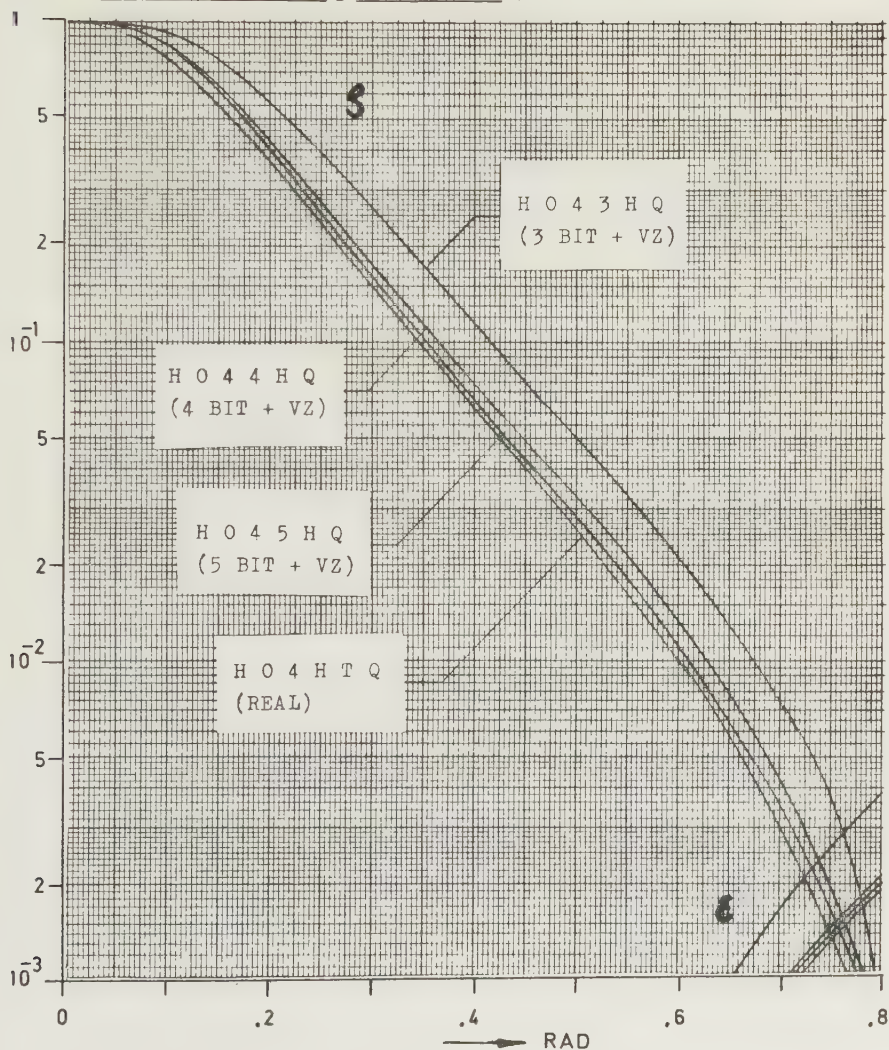


Bild 80. Zurückweisungsrate ξ und Fehlerrate ϵ der Klassifikatoren HO4HTQ, HO45HQ, HO44HQ, HO43HQ, abgetragen über dem Betrag RAD des Fehlervektors. Das Schätzpolynom war bei allen Klassifikatoren in 40 Merkmalen vollständig quadratisch ($L = 861$).

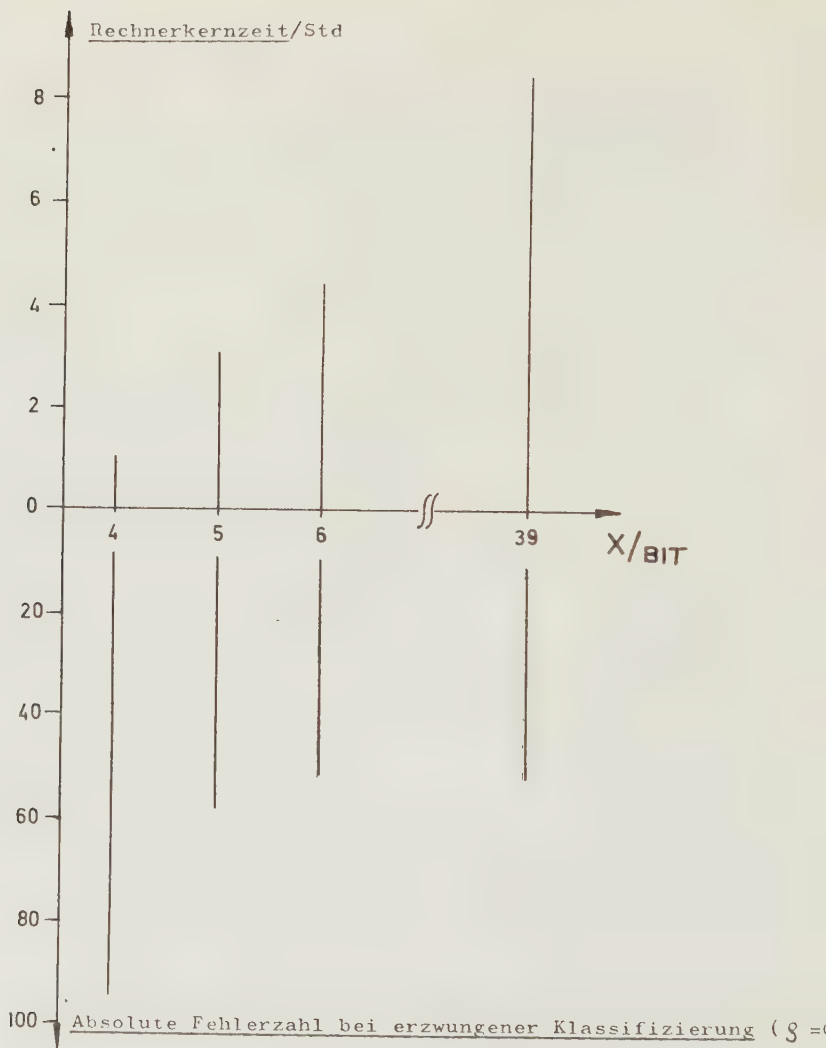


Bild 81. Rechnerkernzeit, die zur Berechnung der Momentenmatrizen der Klassifikatoren HO4HTQ, HO45HQ, HO44HQ, HO43HQ notwendig ist, sowie absolute Fehlerzahl bei erzwungener Klassifizierung, abgetragen über der Zahl X der BITS, die zur Quantisierung der Merkmale w_i verwendet wurden (39 BIT = FIN-REAL-Darstellung).

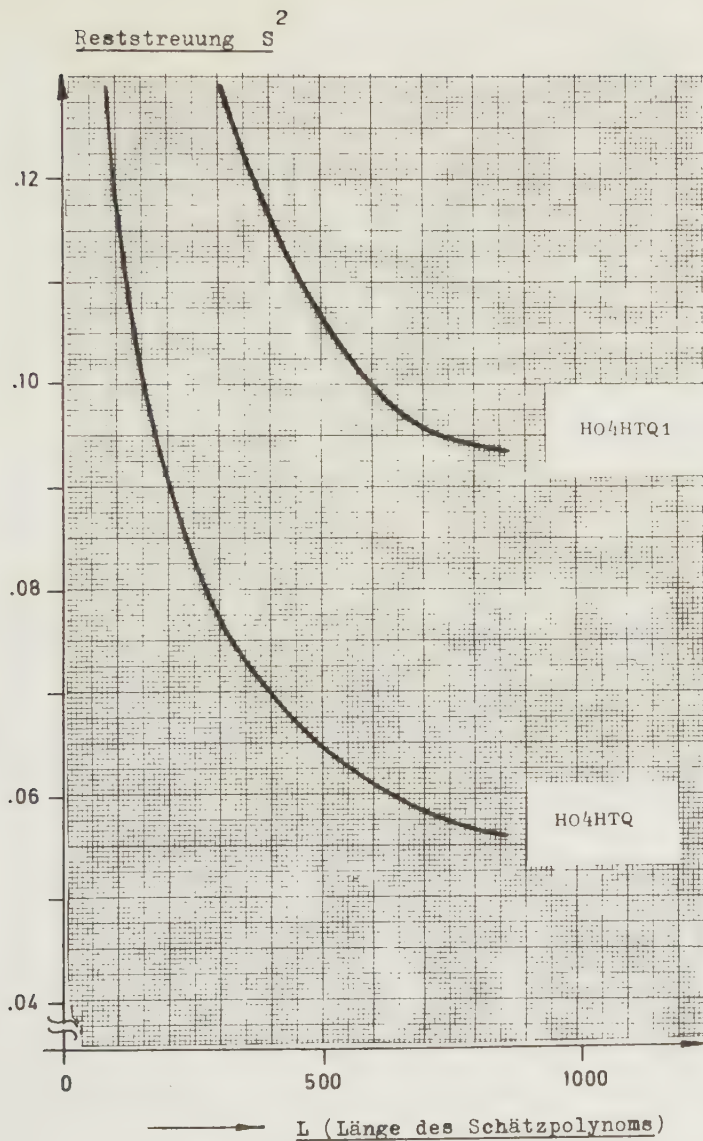
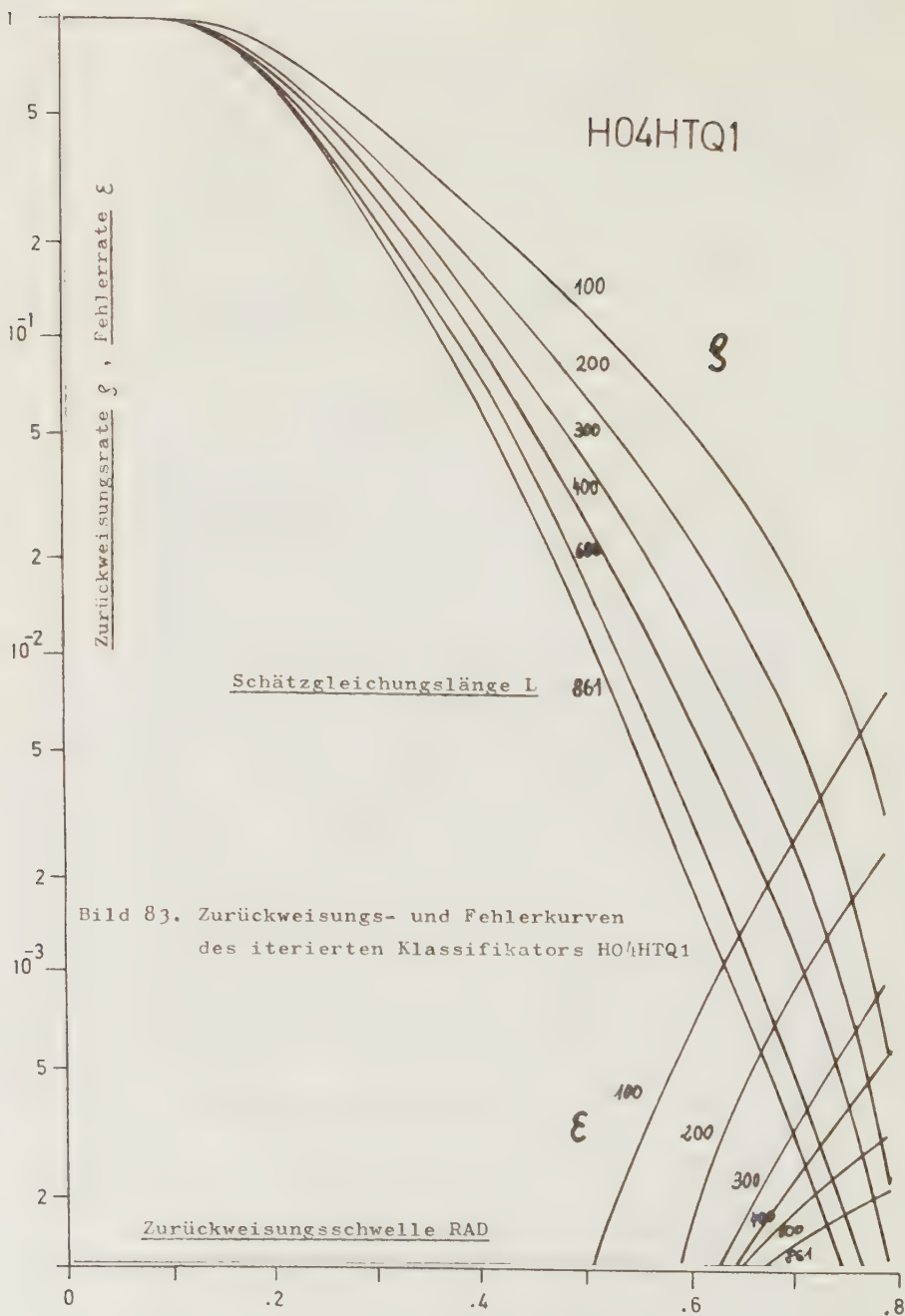
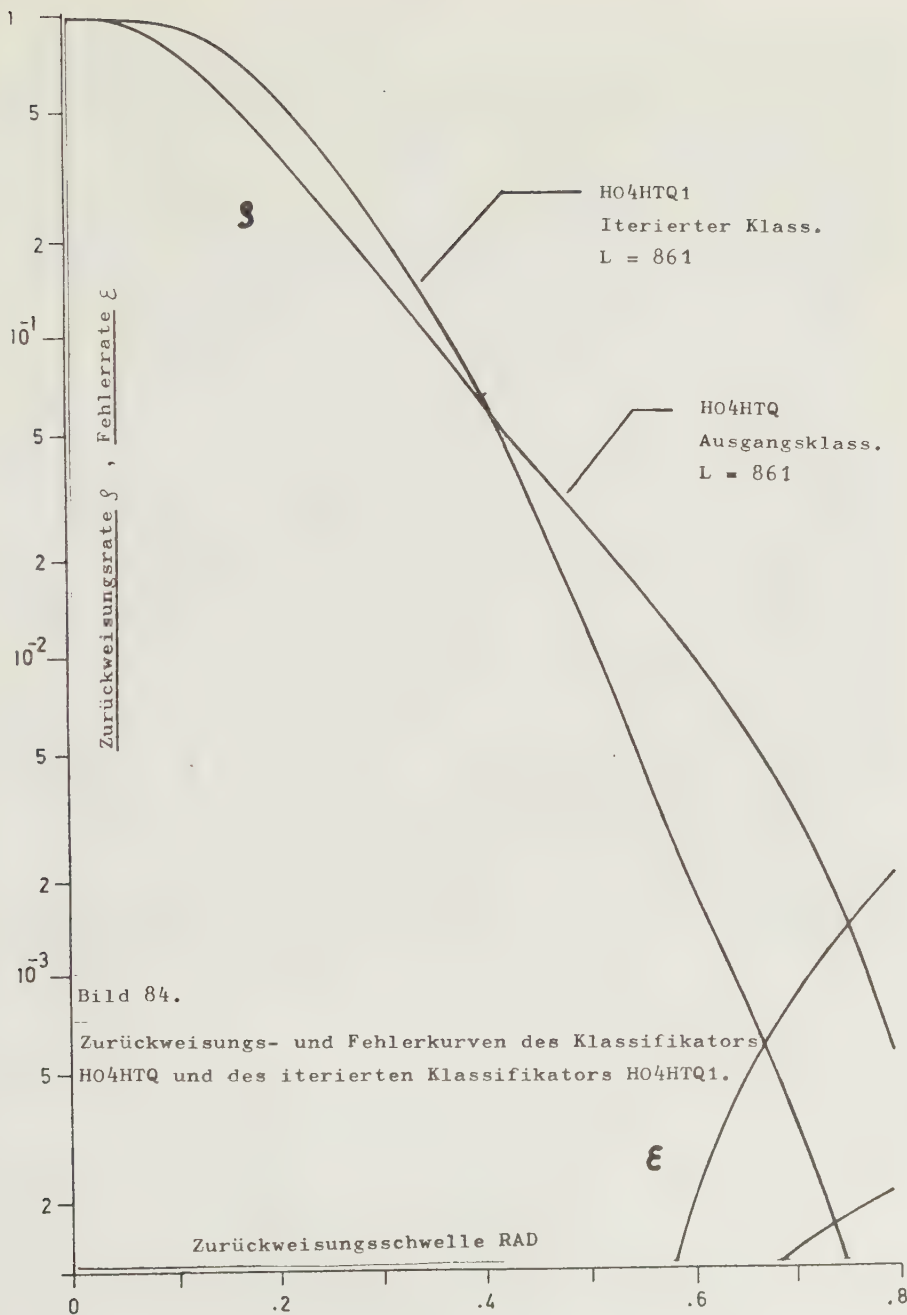


Bild 82. Reststreuung S^2 des Klassifikators HO4HTQ sowie des iterierten Klassifikators HO4HTQ1, abgetragen über der Länge L des Schätzpolynoms.





Absolute Fehlerzahl ohne Rückweisungen bei 22370 Zeichen

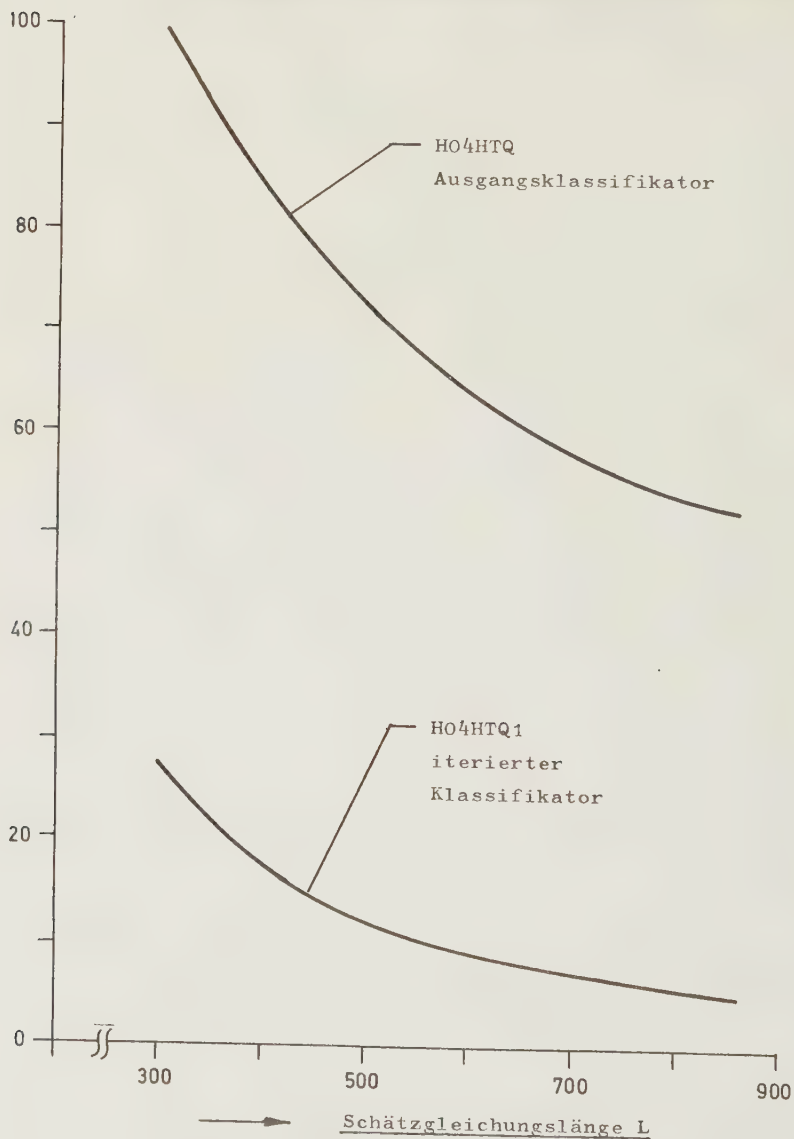


Bild 85. Absolute Fehlerzahlen von Ausgangsklassifikator und iteriertem Klassifikator.

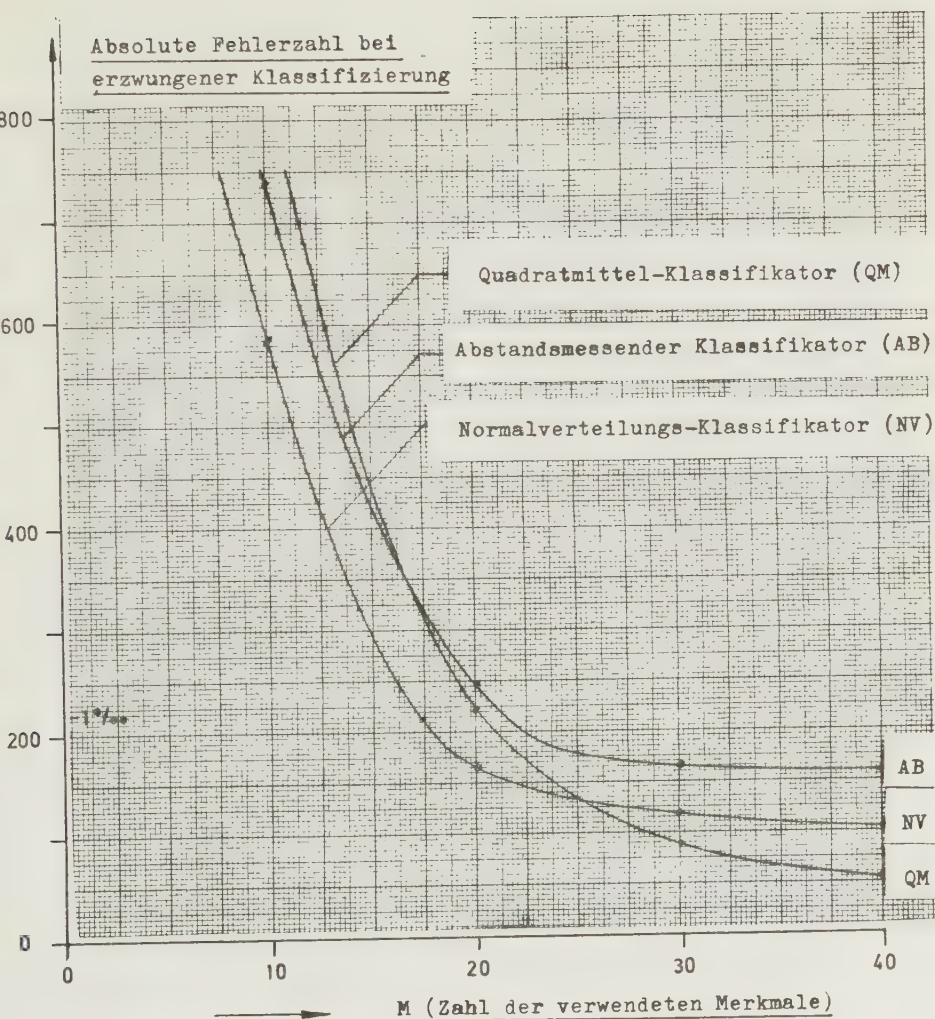


Bild 86.

Absolute Fehlerzahl (bei erzwungener Klassifizierung) von Quadratmittelklassifikator, abstandsmessendem Klassifikator und Normalverteilungsklassifikator, abgetragen über der Zahl M der verwendeten Merkmale.

Stichprobe H0, 22 000 Zeichen.

Zurückweisungsrate bei $\xi=10^{-3}$

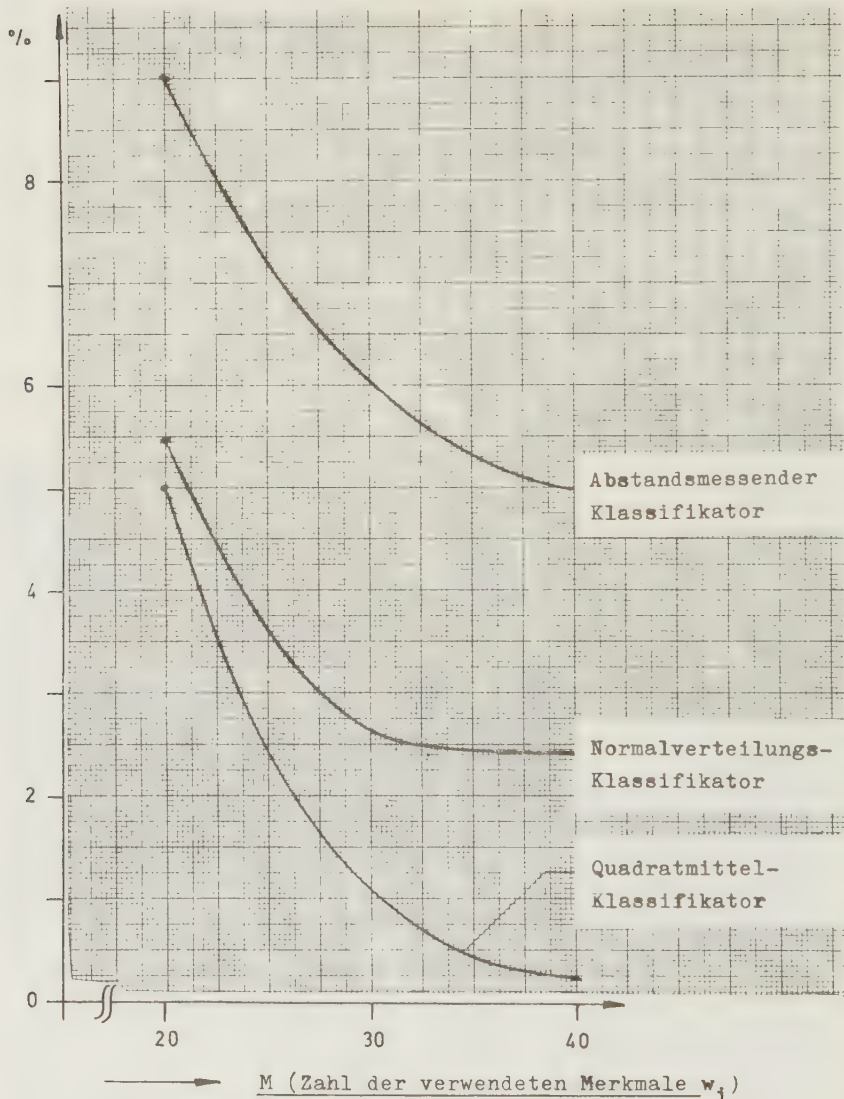


Bild 87. Zurückweisungsrate ξ , die notwendig ist, um eine Fehlerrate von 10^{-3} zu erreichen, abgetragen über der Zahl M der verwendeten Merkmale.

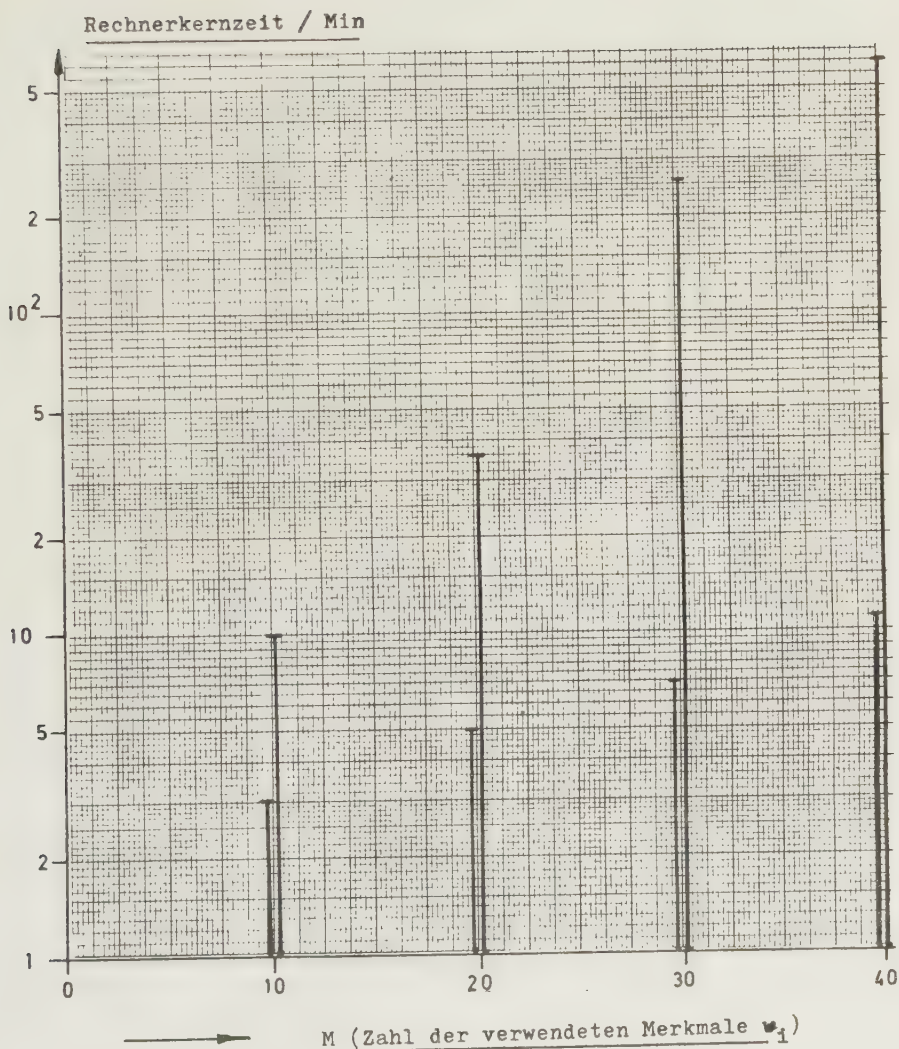


Bild 88. Rechnerkernzeit, die zur Berechnung von Normalverteilungsklassifikator und abstandsmessendem Klassifikator (für diese beiden gilt linker Balken) und Quadratmittelklassifikator (rechter Balken) notwendig ist, abgetragen über der Zahl der verwendeten Merkmale w_i .

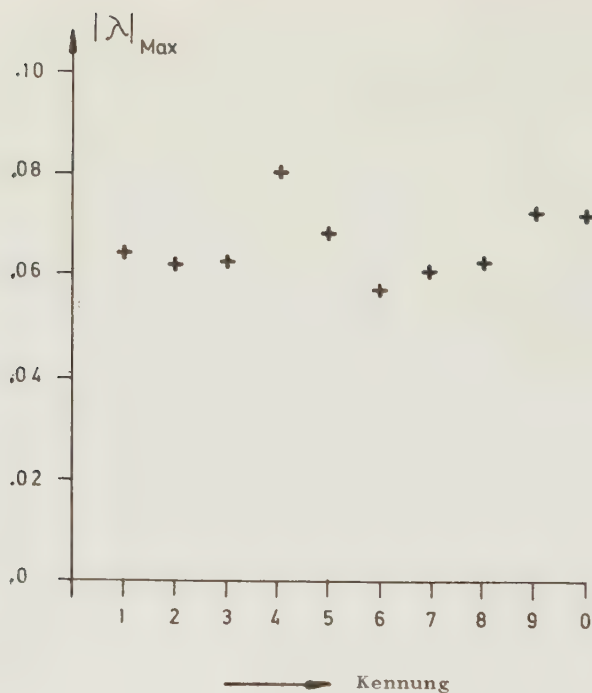


Bild 89. Die betragsgrößten Eigenwerte der klassenspezifischen Kernmatrizen G_k , abgetragen über dem Klassenindex. G_k wurde aus der Koeffizientenmatrix des Quadratmittelklassifikators HO4HTQ berechnet.

KENNUNG	ZAHL DER NEGATIVEN EIGENWERTE
1	24
2	20
3	20
4	18
5	19
6	19
7	15
8	14
9	19
0	19

Bild 90. Tabelle über die Zahl der negativen Eigenwerte der klassenspezifischen Kernmatrix G_k in Abhängigkeit vom Klassenindex.

KENNUNG	KONSTANTE GK

1	-0.145
2	0.102
3	-0.041
4	0.024
5	0.035
6	0.061
7	0.007
8	0.585
9	0.144
0	-0.004

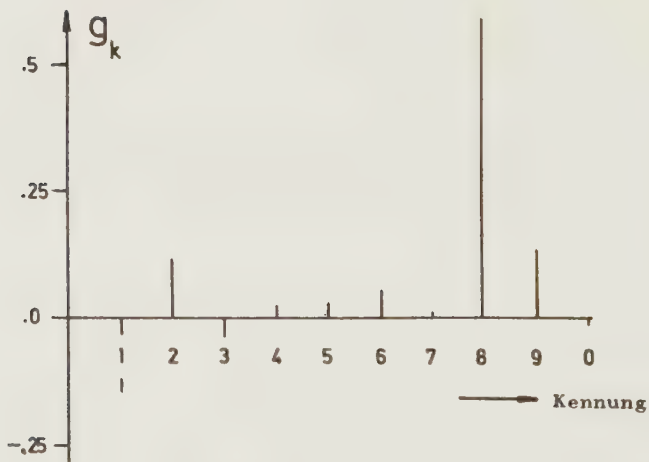


Bild 91. Auflistung der Konstanten g_k (Gleichung (97)), die sich beim Quadratmittelklassifikator H04HTQ ergaben.

An dieser Stelle möchte ich mich recht herzlich bei Herrn Dr. Schürmann und Herrn Prof. Niemann für die vielen Anregungen bedanken, die sich aus zahlreichen Diskussionen ergaben. Mein Dank gilt auch Herrn Prof. Meyer-Brötz, der für meine Arbeit großes Interesse entgegenbrachte, ebenso Herrn Eisenhardt für seine beratende Unterstützung in vielen Fragen der Programmierung. Der Institutsleitung danke ich dafür, daß ich die Möglichkeit hatte, diese Arbeit am Forschungsinstitut in Ulm durchzuführen.

Lebenslauf

Geburtstag-und -ort : 9. September 1945, Esslingen

Eltern : Alfred Ott, Industriekaufmann
Erika Ott, geb. Zaiss

1952 - 1956 : Besuch der Grundschule in Esslingen
(Burgschule, Silcherschule)

1956 - 1965 : Besuch des Schelztorgymnasiums
Esslingen. Schulabschluß: Abitur

1965 - 1966 : Wehrdienst

1966 - 1972 : Studium an der Universität Stuttgart
Studienfach: Elektrotechnik
Fachrichtung: Elektrische
Nachrichtentechnik
Abschluß: Diplom

während verschiedener Semesterferien : Praktikum bei den Firmen Hirschmann
und Siemens

Seit 1972 : wissenschaftlicher Mitarbeiter am
Forschungsinstitut von AEG-Telefunken
in Ulm

